Performance Modelling of Network Management Schemes for Mobile Wireless Networks

Hao Wang

A thesis submitted for the degree of Doctor of Philosophy.
The University of Edinburgh.
September 2009

Abstract

With the rapid development of wireless technologies and user devices, mobile wireless networks have significantly changed people's daily lives in the last two decades. More excitingly, if mobile wireless networks can support Internet Protocol (IP) at their network layers, they will take advantage of the ubiquitously installed IP infrastructure and benefit from the IP-based applications and services developed for wired networks. However, since IP was not designed with consideration of user mobility, this envisioned success of mobile wireless networks will rely heavily on the degree of seamless mobility support and high-quality service provisioning in mobile environments. To meet these requirements, various network management schemes addressing mobility and quality of service (QoS) management in mobile wireless networks have been designed.

In order to examine the extent to which the designed schemes attain their design objectives, i.e. their efficiency and effectiveness, this thesis focuses on the performance modelling of network management schemes for mobile wireless networks. Instead of traditional evaluation approaches such as simulation and queueing theory, a formal performance modelling formalism, named Performance Evaluation Process Algebra (PEPA), is used to conduct the modelling and assessment. PEPA is adopted because of its parsimony and expressiveness of concurrency and compositionality. The PEPA models built in this thesis have generality in that they support high abstractions of the investigated schemes and are independent of detailed implementations. Their structures and behaviour clearly and accurately capture the characteristics of the modelled schemes.

Two important issues in mobile wireless networks are investigated in this thesis. The first issue is about how, in mobile environments, to deploy the Resource reSer-Vation Protocol (RSVP), which is widely used to achieve guaranteed QoS in wired networks. Techniques for solving incompatibilities between RSVP and user mobility are studied. First of all, schemes that simplify the signalling procedure of RSVP are modelled, and their advantages in reducing handover interruptions are verified. In addition, schemes used by a mobile node to make advance resource reservation (ARR) based on RSVP are also studied, and a reservation optimised ARR scheme is proposed. Evaluation results show that the proposed scheme achieves a better network resource utilisation and effectively balances different types of reservation paths in a network, at the reasonable expense of introducing possible service interruptions to slow mobile nodes. The second issue concerns the network selection strategies (NSSs) that are involved in handover management in heterogeneous mobile wireless networks. NSSs have been designed for users to select appropriate networks. To find out the effect of NSSs on both mobile nodes and networks, a general performance evaluation framework, which has an interface to the NSSs used by the mobile nodes and an interface to the resource consumption models of the networks, is investigated. Commonly used NSSs are evaluated from the perspectives of average throughput, handover rate, and network blocking probability. Results of the evaluation explore the effect of these NSSs and their characteristics in different mobility and traffic patterns.

Declaration of originality

I hereby declare that the research recorded in this thesis and the thesis itself was composed and originated entirely by myself in the School of Engineering at The University of Edinburgh.

Hao WANG

Acknowledgements

First and foremost, I wish to thank my supervisors Dr. David I. Laurenson and Prof. Jane Hillston for their invaluable guidance, insights and assistance during the course of my Ph.D. research. I know how to undertake research very much better than I did three years ago. This thesis could not have been possible without their help and support.

The financial support of Mobile VCE, without which I would not have been in a position to commence and complete this work, is also acknowledged and appreciated. I owe thanks to the anonymous reviewers for my first publication, their acceptance of my work was very important to me since it gave me confidence and greatly encouraged me to carry on with my research. I would also like to thank my friends, especially Yan Yan, with whom I spent a wonderful time in Edinburgh.

Special thanks are due to my family who have given me constant support and encouragement. No words can express my deepest appreciation to them and especially to my parents and my sister. I still remember when I was eight years old, at one night I suddenly realised that I had always disappointed my parents for doing poorly in every exam. I whimpered overnight and promised my parents that I would study hard since then. However, I was too young to keep my promise and played Nintendo with my friends on the very next day. Ten years later, when I achieved a high score in the national university entrance examination, for the first time I thought I made my parents proud. Another ten years has passed, I hope that this thesis will make my parents proud of me again, and this thesis is dedicated to them.

To my parents

Contents

List of figures

List of tables

Acronyms and abbreviations

1G	first-generation
2G	second-generation
2D-CTMC	two-dimensional continuous-time Markov chain
3G	third-generation
3GPP	3rd generation partnership project
3GRAN	3G RAN
3GRAT	3G RAT
ABC	always best connected
AP	access point
AR	access router
ARR	advance resource reservation
CAC	call admission control
CCS	Calculus of Communicating Systems
CDMA	code division multiple access
CoA	care-of address
CoV	coefficient of variation
CSP	Communicating Sequential Processes
CTMC	continuous-time Markov chain
DiffServ	differentiated services
DSCP	DiffServ codepoint
EA	enhanced agent

EMPA	Extended Markovian Process Algebra
GPRS	general packet radio service
GSM	global system for mobile communications
GSPN	generalised stochastic Petri net
HHO	horizontal handover
HMIPv6	Hierarchical Mobile IPv6
HMWN	heterogeneous mobile wireless network
HoA	home address
IEEE	Institute of Electrical and Electronic Engineers
IntServ	integrated services
IP	Internet Protocol
IS-95	interim standard 95
ITU	international telecommunications union
ISM	industrial scientific medicine
kbps	kilobit per second
LAN	local area network
LCoA	On-link Care-of-Address
LRD	long-range dependence
MADM	multiple attribute decision making
MAP	Mobility Anchor Point
Mbps	megabit per second
MIPv4	Mobile IPv4
MIPv6	Mobile IPv6

MRS	Markov reward structure
NHO	no handover
NRT	non-real time
NSS	network selection strategy
ODE	ordinary differential equation
PEPA	Performance Evaluation Process Algebra
PHB	per-hop behaviour
QoS	quality of service
RAN	radio access network
RAT	radio access technology
RCoA	regional care-of address
RRSS	relative received signal strength
RSVP	Resource reSerVation Protocol
RT	real-time
SAW	simple additive weighting
SINR	signal to interference and noise ratio
SLA	service level agreement
SMR	session-to-mobility ratio
SMS	short messaging service
SOS	structural operational semantics
SPA	stochastic process algebra
SPN	stochastic Petri net
TIPP	TImed Process for Performance Evaluation

TOPSIS	technique for order preference by similarity to ideal solution
ToS	type of service
VoIP	voice over IP
VHO	vertical handover
W-CDMA	wideband CDMA
WLAN	wireless local area network
WPAN	wireless personal area network
WWAN	wireless wide area network
WRAN	WLAN RAN
WRAT	WLAN RAT

Chapter 1
Introduction

Today, mobile wireless networks have become an integral part of people's daily lives, allowing ubiquitous communications and data transfer. In the near future, mobile wireless networks will be based on Internet Protocol (IP) technologies and people will enjoy more advanced applications and services. Network management schemes in this environment are necessary in order to provide seamless and high-quality mobile services. Before the practical deployment of these schemes, performance evaluation is required to examine their efficiency and effectiveness. This thesis addresses the assessment of the network management schemes by means of developing performance models using a formal technique. The constructed models capture characteristics of the investigated schemes and provide meaningful evaluation results.

1.1 Mobile Wireless Networks

A mobile wireless network generally refers to a network of movable devices that use radio for communications. Due to the nature of radio propagation, wireless systems covering small geographical areas typically provide higher data rates than those that have larger coverage areas. Therefore, mobile wireless networks are usually grouped according to their scales of coverage. The three most influential mobile wireless networks are wireless personal area networks (WPANs), wireless local area networks (WLANs) and wireless wide area networks (WWANs) [1].

1.1.1 Wireless Personal Area Networks

A WPAN covers a small area and is used to connect nearby devices. The most widely used technology for WPANs is Bluetooth, which has been standardised by both the Bluetooth special interest group [2] and the Institute of Electrical and Electronic Engineers (IEEE) 802.15.1 working group [3]. Bluetooth specifies how to construct a low-cost, short-range and general-purpose wireless network using the unlicensed 2.4-

GHz industrial scientific medicine (ISM) band. It has achieved global acceptance such that any Bluetooth enabled device can connect to others and form a Bluetooth network.

Bluetooth 2.1, which was ratified in 2007, supports a gross data rate of 1 megabit per second (Mbps) in basic rate mode or, in enhanced data rate mode, a gross data rate of 2 or 3 Mbps. The latest Bluetooth 3.0 ratified in 2009 can even provide a throughput of data at the approximate rate of 24 Mbps. The fundamental form of communication in a Bluetooth network is a *piconet*, which consists of a master node and up to seven active slave nodes[I] within a range of about 10m.[II] In the case of multiple active slaves, the network topology is referred to as point-to-multipoint. That is, all communications within the network only happen between the master node and a slave node, and direct communications between slave nodes are not allowed. Two or more piconets can be interconnected and form a larger network, called a *scatternet*. Each piconet in this scatternet has its own master node, and the master node of a piconet could only be a slave node in other piconets. Communications between two piconets can be established if at least one node from each piconet is within the reach of another and they form a master-slave relationship.

Unlike other communication technologies that focus primarily on the physical, data link, and possibly networking aspects of communications, the Bluetooth technology also specifies precise *profiles* that support a wide range of applications. For example, the *headset* profile describes how a Bluetooth enabled headset should communicate with a Bluetooth enabled device such as a computer or a mobile phone. The *personal area networking* profile describes how two or more Bluetooth enabled devices can form an ad hoc network and how the same mechanism can be used to access a remote network (e.g. the Internet) through a network access point (AP).[III] With the Bluetooth technology, users can move data files of videos, music and photos between their own devices and trusted devices of others; information about calendars and addresses can be synchronised between devices.

[I] There can be up to 255 further inactive slave nodes. They are called parked nodes and can be brought into active state by the master node.

[II] Depending on its maximum radio power, i.e. 1mW, 2.5mw and 100mW, a Bluetooth device has a coverage range of 1m, 10m and 100m respectively. 10m is the most typical range.

[III] An AP offers wireless link connection and functions as a relay point between a wired network and a wireless network.

1.1.2 Wireless Local Area Networks

The concept of WLAN was developed with the aim of building a local area network (LAN) in which connections between nodes are wireless links rather than cables. The most widely adopted WLAN standard is IEEE 802.11, which consists of a family of standards that defines the physical layer and the medium access control layer of a WLAN. The legacy IEEE 802.11 standard [4], which was ratified in 1997 and clarified in 1999, features a data rate of 1 or 2 Mbps using both radio transmission in the unlicensed 2.4-GHz ISM band and infrared transmission. On this basis, a family of amendments to the original specification have been standardised and some well-known ones are:

- IEEE 802.11*a*: operates in the 5-GHz ISM band and allows a maximum throughput of 54 Mbps. The data rate can be reduced to 48, 36, 24, 18, 12, 9 then 6 Mbps if required.
- IEEE 802.11*b*: operates in the 2.4-GHz ISM band as the legacy standard and enhances it to support data rates of 5.5 and 11 Mbps, besides the basic 1 and 2 Mbps.
- IEEE 802.11*g*: also operates in the 2.4-GHz ISM band and achieves the same group of data rates as that of IEEE 802.11*a*. Moreover, it is fully backwards compatible with IEEE 802.11*b* and supports the data rates provided by IEEE 802.11*b* as well.
- IEEE 802.11*e*: defines mechanisms to support delay-sensitive applications, such as voice over IP (VoIP) and streaming multimedia.
- IEEE 802.11*i*: specifies a framework for providing security management.

There are two basic operational modes in an IEEE 802.11 WLAN: *infrastructure* mode, and *ad hoc* mode. Within the infrastructure mode, a WLAN consists of at least an AP which is usually connected to a fixed network infrastructure (e.g. the Internet) and a number of IEEE 802.11 enabled nodes that associate with the AP. The nodes within and outside the WLAN communicate via the AP. The ad hoc mode simply represents a group of IEEE 802.11 enabled nodes that communicate directly with each other without an AP and a connection with a fixed network.

The geographical coverage of a WLAN in an indoor environment is usually about several dozens of meters, and that in a outdoor environment is normally about one or two hundred metres [1]. WLANs have gained unexpected market penetration due to their flexibility, high data rates, low cost of deployment and use. They are now widely used in companies and homes to provide wireless data services inside buildings and are also being deployed to provide network access in public areas where there are a high density of users such as airports and coffee shops.

1.1.3 Wireless Wide Area Networks

WWANs, usually referred to as mobile cellular networks, can provide public mobile services over large geographical areas to users moving at both pedestrian and vehicular speeds. WWANs are generally classified into *generations* based on the technologies they use. In the early 1980s, first-generation (1G) WWANs using analog radio technologies became commercially available. They were mainly used for voice, although data communication was also supported at the maximum user data rate of 1.2 kilobit per second (kbps) [5]. At this slow data transmission rate, only a very few data applications can be used, for example paging.

The major step in WWAN evolution was the introduction of digital modulation and digital signal processing technologies, which led to second-generation (2G) WWANs in the early 1990s [6]. Besides making voice calls, 2G WWANs also provide services like call diverting, setting up closed user groups in which internal calls are charged at reduced rate, and the short messaging service (SMS) that enables users to transmit alphanumeric pages of limited length (160 7-bit characters). The two most popular 2G standards used throughout the world are global system for mobile communications (GSM) and interim standard 95 (IS-95). GSM provides a maximum user data rate of 9.6 kbps and is widely deployed in Europe and parts of Asia. IS-95 is also known as cdmaOne due to its usage of code division multiple access (CDMA) technology. By using CDMA, a user's traffic can be transported over a frequency band which is much broader than the spectrum occupied by the user's original traffic (1.25-MHz in IS-95 and 200-kHz in GSM). IS-95 supports a maximum user data rate of 9.6 or 14.4 kbps, and is a popular choice in North America and Korea. As the requirements of mobile data services grow, 2G WWANs have been upgraded into what are commonly referred to as 2.5G WWANs with higher data rates. For example, general packet radio service (GPRS) is an enhancement of the GSM standard and is a near-term solution

for the delivery of Internet packet data to users. GPRS has a peak data throughput of 171.2 kbps and is well suited to non-real time Internet usage such as asymmetric web browsing where users download much more data than they upload. The 2.5G enhancement for IS-95 is called IS-95B and its data rate can reach 64 kbps.

In the late 1990s, standardisation efforts for third-generation (3G) WWANs were carried out featuring high data rates, seamless integration of cellular networks with the Internet, and high-quality multimedia applications. Although the international telecommunications union (ITU) formulated a plan to implement a single, ubiquitous wireless communication standard for all countries throughout the world, the eventual 3G evolution remains split between two major technologies: wideband CDMA (W-CDMA) and cdma-2000. W-CDMA is based on the network fundamentals of GSM. It requires a minimum bandwidth of 5-MHz and supports data rates up to 2 Mbps, thereby allowing multimedia broadcasting, streaming audio and video, and interactive Internet games. Because W-CDMA requires expensive hardware equipment upgrade, the installation of W-CDMA is likely to be slow. On the other hand, cdma-2000 gives a less expensive upgrade path by maintaining seamless backward compatibility with cdmaOne and IS-95B user equipment and using the original 2G CDMA as the building block. The first cdma-2000 air interface uses a single cdmaOne radio channel and yields typical throughput rates of up to 144 kbps. Later cdma-2000 allows simultaneous usage of three adjacent or non-adjacent radio channels, and the peak data rate in this case is in excess of 2 Mbps.

1.2 Service Provisioning in Mobile Wireless Networks

As presented in the previous section, there has been a trend in mobile wireless networks towards increasing bandwidth, thus facilitating access to information, applications and services available over the Internet. To achieve this, mobile wireless networks are required to support IP, which is a universal network layer protocol for the Internet, for data packet delivery so that they could easily provide IP-based data and multimedia services to mobile users. However, since the Internet was not designed with the awareness of node mobility, this envisioned success of mobile wireless networks will rely heavily on the degree of seamless and high-quality service provisioning in mobile environments. To meet these requirements, network management schemes addressing mobility and quality of service (QoS) management

in mobile wireless networks have been designed. Before the practical deployment of these schemes, it is necessary and important to examine to what extent they achieve their design objectives. Therefore, the primary motivation of this thesis is the performance modelling of the network management schemes that facilitate seamless and high-quality services in mobile environments in order to assess their efficiency and effectiveness, under the requirements that the performance models should have clear and accurate representations of the mechanisms underlying the modelled schemes, whilst maintaining model generality in order that they are independent of detailed implementations. To achieve these aims, a formal performance modelling formalism, named Performance Evaluation Process Algebra (PEPA), is adopted to provide the modelling and assessment.

Two important issues regarding service provisioning in mobile wireless networks are discussed in this thesis. The first issue is how to provide QoS to mobile users. The term QoS is used in many meanings ranging from a user's perception of a service to a set of connection parameters necessary to achieve particular service quality. The best-effort data packet delivery, which is the basic QoS offered by the Internet, cannot satisfy the requirements of multimedia and real-time traffic. Therefore, QoS architectures such as integrated services (IntServ) and differentiated services (DiffServ) have been proposed to augment this basic Internet service model with various service classes suitable for different applications and scenarios. However, these QoS architectures were initially designed for wired networks, and when they are deployed in mobile environments, service interruptions could happen due to user mobility. For example, in mobile wireless networks supporting the IntServ architecture, which provides end-to-end guaranteed QoS by reserving resources along the path between communicating ends, a mobile node has to request a new reservation path after it changes its network point of attachment. There are several problems result from this reservation re-establishment procedure and network management schemes have been designed in order to make this procedure smoother and reduce overheads. These network management schemes will be discussed and evaluated in Chapter 4 and Chapter 5.

The second issue is about the future all-IP network that integrates heterogeneous wireless networks such as those introduced in Section 1.1. This heterogeneous environment provides mobile users with the potential for taking advantage of different

access technologies. A mobile node equipped with different radio access interfaces can choose an appropriate network to access its services, depending on a number of different aspects such as user preferences, device capacity, application demands, etc. Therefore, the selection mechanism used by the mobile node determines which network a mobile node will connect to and controls the session behaviour of the mobile node. Consequently, it is important and meaningful to investigate how it affects both the performance of user applications at the mobile node end and the utilisation of network resources. The network selection mechanisms will be discussed and evaluated in Chapter 6.

1.3 Contributions

The primary contribution of the work presented in this thesis is that it carries out the first investigation on formal performance models of network management schemes for mobile wireless networks. More precisely, the contributions of this thesis can be summarised as follows:

- Performance models of schemes that have been designed for solving incompatibilities between an important resource reservation protocol used for attaining guaranteed QoS and user mobility are built. These models fulfil the above modelling requirements, verify the rationality and effectiveness of the designed schemes, and yield clear explanations as to how seamless QoS provisioning is achieved in mobile environments by means of these schemes.

- As one of the solutions to mobility and QoS integration, a novel advance resource reservation scheme which improves network resource utilisation is proposed. This scheme is designed in a modular way and consists of two admission control mechanisms that can be easily integrated into existing schemes. It discriminates resources that are actively and passively reserved by mobile nodes and allocates them appropriately. The proposed scheme achieves an enhancement of the QoS perceived by mobile users.

- A general performance evaluation framework for strategies used by mobile nodes to select appropriate networks in heterogeneous wireless network environments is constructed. This framework gains its generality by having an interface to models capturing network resource consumption and an interface

to different strategies used by mobile nodes, in the forms of network blocking probability and network selection probability respectively. Not only the effect of various strategies on the mobile nodes and the networks but also their characteristics in different mobility and traffic scenarios are explored by this framework.

- The above framework also provides a novel approach to the derivation of network blocking probability and handover rate in heterogeneous mobile wireless networks when different selection strategies are used. This approach employs an iterative algorithm which links models of network resource consumption and models of different selection strategies by interchanging necessary parameters between them. Its convergence speed is fast and performance measures of interest are direct results of the algorithm.

1.4 Organisation

This chapter presents a basic introduction to the most well-known mobile wireless networks, and highlights the motivations and contributions of this thesis. The reminder of this thesis consists of the following chapters:

Chapter 2 continues the introduction of the preliminaries related to the network management schemes that are modelled in this thesis. These include mobility management protocols which solve the problem of continuous delivery of packets to mobile nodes; the architectures designed for providing QoS in wired networks; and handover management in heterogeneous wireless network environments.

Chapter 3 initially reviews traditional performance modelling methods, followed by an introduction to the syntax and semantics of the performance modelling formalism, PEPA. An example is used to demonstrate how PEPA can be used for system performance evaluation.

Chapter 4 studies one of the two major problems associated with a widely used resource reservation protocol for QoS provisioning when it is deployed in mobile environments: how to optimise its signalling procedure in order that interruptions and overheads to mobile nodes can be reduced. PEPA models of

both basic and optimised signalling procedures are built, and the advantages of procedure optimisation are verified.

Chapter 5 addresses the other problem of resource reservation in mobile wireless networks: how to efficiently reserve resource in advance for mobile nodes. A scheme which optimises existing solutions in order to improve network resource utilisation is proposed. The benefits of the proposed scheme is demonstrated through performance comparison between PEPA models of different schemes.

Chapter 6 considers a particularly popular heterogeneous wireless network environment: 3G and WLAN interworking networks. The focus is put on strategies used by mobile nodes to choose appropriate networks. A general performance evaluation framework, which explores the effect of the strategies and their characteristics, is investigated.

Chapter 7 concludes the thesis by reviewing the results of the previous chapters. The limitations of this thesis and possible directions of future work are also indicated.

Chapter 2
Network Management Schemes for Mobile Wireless Networks

2.1 Introduction

Owing to the unprecedented growth of wireless communications, the Internet has developed from the traditional interconnected wired networks to interworking wired and wireless networks, and will eventually evolve towards an all-IP network if all the wired and wireless networks support IP at the network layer. Mobile wireless networks in this all-IP network are envisioned to become more robust and cost effective, because they can take advantage of the ubiquitous installed IP infrastructure and thus benefit from IP-based applications and services developed for wired networks. However, since IP was not designed with consideration of user mobility, this envisioned success of mobile wireless networks will rely heavily on the degree of seamless mobility support and high-quality service provisioning. Several issues need to be considered before mobile wireless networks are integrated into the all-IP network.

The first issue is that a mobile node must be able to remain attached to the Internet when it changes its point of attachment from one network or subnet to another. This is not a easy task because an IP address of a node is used as both its identifier by transport or higher layer protocols and its network layer locator for receiving packets. Therefore, a mobile node will have no mobility if it just uses a static IP address for its communications. Another issue associated with node mobility is how to provide QoS in mobile wireless networks. Like IP, QoS management architectures were initially designed for stationary nodes. When they are directly deployed in mobile environments, node mobility can cause service interruptions to mobile nodes. Moreover, the future all-IP network will integrate heterogeneous wireless networks such as 3G cellular networks, WLANs and Bluetooth. A mobile node may be equipped with different radio access interfaces and receive services through alternative wireless networks. In these heterogeneous mobile wireless networks

(HMWNs), only providing seamless mobility is not enough. The mobile node may want to be always best connected (ABC) [7]. That is, the mobile node not only wants to be always connected, but also wants to be connected to the best possible network. To achieve the ABC requirement, a handover management technique must choose an appropriate wireless network for a specific service.

This chapter presents the preliminaries that are related to the network management schemes which are modelled in this thesis. Section 2.2 introduces the mobility management protocols which solve the problem of delivering packets to mobile nodes. The architectures designed for QoS provisioning in wired networks and their inefficiency in mobile environments are described in Section 2.3. In Section 2.4, the issue of handover management in HMWNs is discussed. Finally, Section 2.5 gives a summary of this chapter.

2.2 Mobility Management Protocols

To provide connectivity to mobile nodes as they move across different networks, many solutions are proposed to enrich IP with the capability of managing the node mobility. The most well-known of such proposals are Mobile IPv4 (MIPv4) [8] and Mobile IPv6 (MIPv6) [9]. Although very effective, they are not efficient enough in all scenarios and some extensions such as Hierarchical Mobile IPv6 (HMIPv6) [10] have been suggested. Based on their respective scopes of applications, mobility management protocols can be divided into two classes: *macro-mobility protocols* and *micro-mobility protocols*.

2.2.1 Macro-mobility Protocols

The MIPv4 and MIPv6 are usually called macro-mobility protocols in literature. In both protocols, a mobile node effectively utilises two IP addresses, one for identification and the other for routing. The former address is called the home address (HoA), which is a long-term IP address on the mobile node's home network[1] and is administrated in the same way as an IP address of a stationary node. The latter address is called the care-of address (CoA), which is associated with the mobile node when it is out of its home network and reflects its current point of attachment.

[1]The home network of a mobile node is the network whose network prefix matches that of the mobile node's HoA.

2.2.1.1 Mobile IPv4

MIPv4 [8] is the mobility support protocol for IPv4 Internet. Each mobile node is always identified by its HoA, regardless of its current location. To support node mobility, some access routers (ARs)[II] are configured as mobility agents with which a mobile node should register. A mobility agent in the mobile node's home network and its visited network is called a *home agent* and a *foreign agent* respectively. The operation of MIPv4 consists of three mechanisms, namely agent discovery, registration, and tunnelling [11].

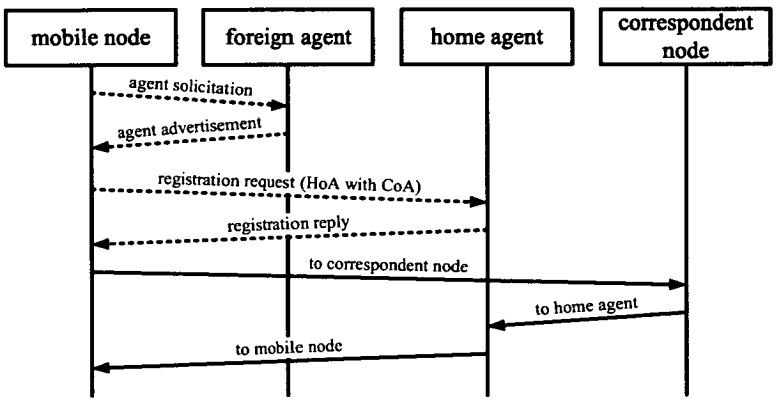


Figure 2.1: *The basic operation of MIPv4*

Figure 2.1 shows the basic operation of MIPv4, in which dashed arrows represent signalling packets and solid arrows represent data packets. A mobility agent makes itself known by sending agent advertisement messages. A mobile node may optionally solicit agent advertisements from any locally attached agent. These agent advertisement messages can include information such as the network prefix of the mobility agent, from which the mobile node can determine whether it has changed its point of attachment. When the mobile node detects that it has moved to a new network, it obtains a CoA[III] on that network and registers this address with its home agent by exchanging registration request and registration reply messages. A successful registration creates or modifies a *mobility binding* at the home agent, associating the mobile node's HoA with its current CoA. The home agent will then be able to intercept packets destined for the mobile node's HoA and tunnel them to the mobile node's

[II] An AR resides on the edge of a network and offers IP connectivity to mobile nodes by providing them with routing services.

[III] There are two different types of CoAs: a foreign agent CoA that is the IP address of the foreign agent; and a co-located CoA that is externally obtained and associated with one of the mobile node's interfaces. The choice depends on the administrative configuration.

CoA. The tunnelling process can be achieved by encapsulation algorithms such as IP-within-IP encapsulation [12]. That is, when the home agent intercepts a data packet, it precedes it with a new IP header whose source address is the address of the home agent and destination address is the CoA of the mobile node. In this way, this packet will be diverted to the CoA and after decapsulation, i.e. removing the outer IP header, the original IP packet is extracted. When the mobile node comes back to its home network, it deregisters with its home agent and only uses its HoA.

As can be seen from Figure 2.1, one problem of MIPv4 is called *triangle routing*. That is, packets destined for a mobile node must travel through its home agent, whereas packets from it can be routed directly to its correspondent node. This problem can be solved by an extension protocol called *route optimisation* [13], in which the correspondent node can be informed of the mobile node's current CoA by the home agent and then communicate directly with the mobile node.

2.2.1.2 Mobile IPv6

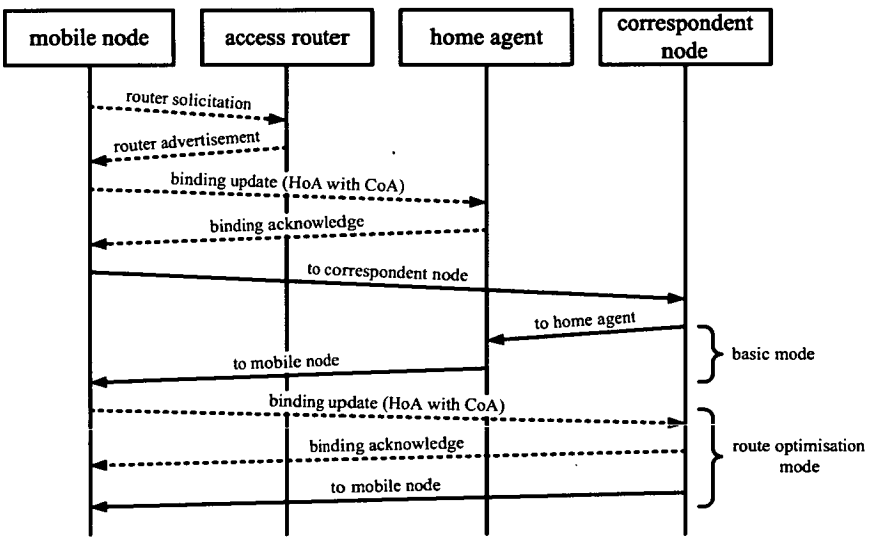


Figure 2.2: *The basic mode and route optimisation mode of MIPv6*

Due to the limited address space in IPv4, IPv6 is designated as the successor to IPv4 and MIPv6 [9] is designed to provide mobility support in the IPv6 Internet. Since MIPv6 utilises the IPv6 version of address autoconfiguration, it does not require the support of foreign agents. The basic operation of MIPv6 is similar to that of MIPv4 and is shown in Figure 2.2. A mobile node detects its movement by IPv6

router advertisements and when it is away from its home network, it registers the binding of its HoA and its current CoA with its home agent. The home agent will intercept the packets destined for the mobile node's HoA and forward them to the corresponding CoA using IPv6 encapsulation [14]. However, unlike in MIPv4 in which the registration process must be carried out by exchanging separate registration messages, in MIPv6 the binding information is included in the *mobility extension header* of a IPv6 packet and can either be sent in separate signalling messages or be included in normal data packets. This design can reduce both the registration latency and the handover signalling cost.

MIPv6 also provides the route optimisation mode, which requires the mobile node to register its binding at its correspondent node. Packets from the correspondent node can be routed directly to the CoA of the mobile node. However, the correspondent node uses an IPv6 *routing header* to route the packet to the mobile node, rather than using encapsulation as the home agent. In this way, the mobile node can determine whether its correspondent node knows its latest CoA. If an encapsulated packet is received, the mobile node will immediately inform the correspondent node of its current location.

2.2.2 Micro-mobility Protocols

As presented in Section 2.2.1, before packets can arrive at the correct address of a mobile node, two procedures must be carried out: movement detection and location registration. During the period of these procedures, packets will be sent to the mobile node's previous CoA and might be dropped. This situation deteriorates if the home agent and the correspondent node are far away from the mobile node in topology. Moreover, if there are many mobile nodes in the network, there will be a large volume of signalling traffic in the network. Therefore, micro-mobility protocols such as HMIPv6 have been designed to reduce the latency and overheads that are associated with the location registration. Some other micro-mobility extensions such as MIPv4 fast handover [15] and MIPv6 fast handover [16] have also been proposed to improve the movement detection procedure of MIPv4 and MIPv6 respectively, but they are not as widely used as HMIPv6. Detailed reviews of micro-mobility protocols can be found in [17, 18].

2.2.2.1 Hierarchical Mobile IPv6

HMIPv6 [10] is a direct extension of MIPv6 by utilising a special mobility agent called Mobility Anchor Point (MAP) to limit the scope and amount of handover signalling. A MAP can be located at any level in a hierarchical network of routers. It usually covers a group of ARs and forms a MAP domain. When a mobile node moves into a new MAP domain[IV], it acquires an On-link Care-of-Address (LCoA) referring to the AR to which it is connected and a regional care-of address (RCoA) referring to its serving MAP. The mobile node first registers the binding of its current RCoA and LCoA with its serving MAP, and then the binding of its HoA and RCoA with its home agent and correspondent node.

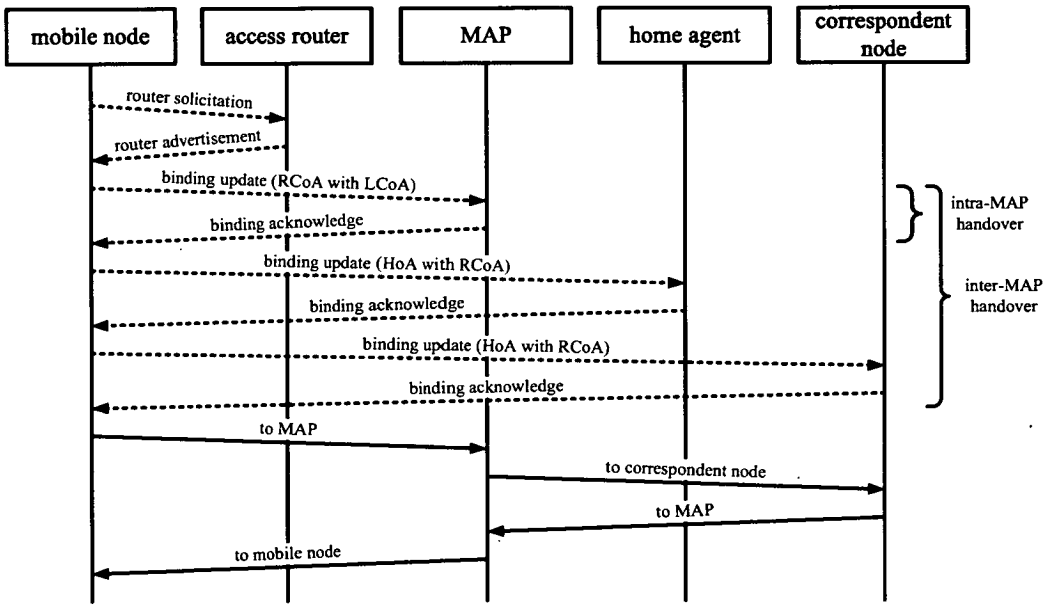


Figure 2.3: *The basic operation of HMIPv6*

Outside the MAP domain, the mobile node is identified by its RCoA. As shown in Figure 2.3, all the packets sent to the mobile node are addressed to its RCoA and they are intercepted by the MAP and forwarded to the mobile node's LCoA using IPv6 encapsulation. That is, when the MAP intercepts a data packet, it precedes it with a new IP header whose source address is the address of the MAP and destination address is the LCoA of the mobile node. In the opposite direction, all the packets sent to the correspondent node must be encapsulated by the mobile node by setting

[IV] Like MIPv6, movement detection is implemented by listening to the router advertisement messages from ARs, which include MAP information.

the source address of the outer header to the LCoA of the mobile node and the destination address to the address of the MAP. The MAP will then decapsulate the packet and forward it to the correspondent node. When the mobile node performs a micro handover, i.e., switches to a different AR within the same MAP domain, it only registers its new LCoA with the MAP, and there is no signalling message outside the MAP domain. Therefore, the movement of the mobile node within the MAP domain is transparent to its home agent and correspondent node. In this way, the handover latency is reduced because the MAP is usually near to the mobile node, and outbound signalling is minimised.

2.3 QoS Management Architectures

Internet traffic generated by multimedia applications and services requires a high level of quality and imposes great demands on the network. To provide users with satisfactory services, traffic-specific requirements of the users on QoS related parameters such as bandwidth, delay and packet loss rate must be fulfilled. The straightforward and simplest approach to achieve this is over-provisioning of network resources. However, over-provisioning is not always feasible for technical and economical reasons [19]. As an alternative, QoS management architectures can be used by network administrators for dynamic and optimum network resource usage. The most well-known QoS management architectures are IntServ [20] and DiffServ [21].

2.3.1 Integrated Services Architecture

The essential idea of the IntServ architecture is to extend the basic best-effort service provided by the Internet, in order that an application can choose its required QoS from a range of different levels of services provided by the network. Two types of services are included in IntServ: *guaranteed service* and *controlled-load service*. The guaranteed service provides hard QoS assurances that the end-to-end packet delay is kept within agreed upper bounds, and also that no packets will be dropped due to router buffer overflow [22]. Accordingly, this type of service is appropriate for real-time applications with strict requirements on packet delivery time. The controlled-load service provides qualitative QoS guarantees that its users would experience a lightly loaded best-effort network [23]. It offers a high probability that packets are successfully delivered to the destination, and a high percentage of packets not greatly

exceeding the minimum delay experienced by any successfully delivered packet. Real-time applications which operate well in lightly loaded networks but degrade badly in the presence of network congestion can use this type of service.

The services provided by the IntServ architecture are *per-flow* based. In the context of IntServ, a flow is a stream of packets with the same source and destination addresses and port numbers. Routers along the data flow path should maintain the QoS requirements of a data flow, called a *state* of that flow, so that end-to-end QoS can be provided. More importantly, the required services cannot be provided unless routers are able to reserve resources such as buffers for a data flow. For that reason, the IntServ architecture employs a reservation setup protocol, Resource reSerVation Protocol (RSVP) [24], to first check for available resources and then create and maintain flow-specific information along the data flow path.

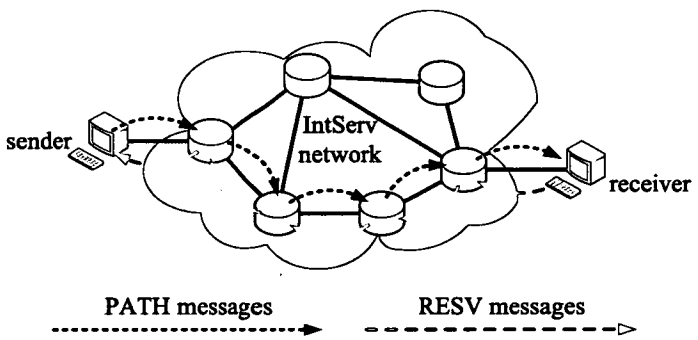


Figure 2.4: *The basic operation of RSVP*

RSVP is a general end-to-end QoS signalling protocol which can be used by a host to request specific services from the network for its data flows. It treats a sender as logically distinct from a receiver and reserves resources in only one direction. Therefore, if a host acts as both a sender and a receiver in the communication with its correspondent node at the same time, two reservation paths between them need to be established. The basic operation of RSVP is shown in Figure 2.4. To start a QoS session, the sender sends a PATH message towards the receiver. The PATH message contains information about the traffic characteristics such as peak packet rate and packet size of the upcoming session. Every router along the path forwards the PATH message according to some routing protocol. If an intermediate router is RSVP-compliant, it installs a state for the session and may optionally add its capability information such as link delay and throughput in the PATH message. Upon receiving the PATH message,

the receiver determines what type of reservation it should make and then responds with a RESV message. The RESV message follows the *reverse route* of the PATH message and requests resources for the session. Each RSVP-compliant router makes its local admission decision based on a comparison of the requested and the available network resources, along with some additional policy control. The actual resource allocation for a flow is implemented by the packet classifier and packet scheduler components in a router. The packet classifier determines what class of service a packet should receive and the packet scheduler manages the output queue according to the service class of a packet. Routers that do not support RSVP simply pass the messages transparently. If all the RSVP-compliant routers along the data flow path accept the reservation request, a reservation path is established between the sender and the receiver. Moreover, reservation paths in RSVP are maintained in a *soft* way. That is, a reservation path has a lifetime associated with it. To keep the reservation path active, the sender must periodically send PATH messages and the receiver must respond accordingly. Otherwise, the reservation path will expire.

The major drawback of the IntServ architecture is its poor scalability, which results from its per-flow traffic handling mechanism [25]. The storage and processing overheads on routers increases proportionally with the number of flows. Moreover, reservation paths need to be regularly refreshed thereby adding traffic on the network and they may time out due to the loss of refreshing packets which are not sent by a reliable transport protocol. Despite the above problems, the IntServ architecture is still viewed as a valuable component in a broader set of QoS technologies because reserving resources is the only way to provide services with guaranteed quality [26].

2.3.2 Differentiated Services Architecture

The DiffServ architecture is designed to cope with the scalability problem faced by IntServ and implement service differentiation in the Internet. This architecture does not use signalling mechanisms, and QoS is applied to aggregated traffic rather than specific flows. In DiffServ, service differentiation is achieved by dividing the traffic into a small number of classes and treating each class differently. DiffServ uses the type of service (ToS) field in the IP header, which is called the DiffServ codepoint (DSCP), for marking a packet with a particular class.

When a packet arrives at an ingress router of a DiffServ domain, it is classified and assigned a DSCP. The DiffServ architecture achieves its scalability by putting

packet classification and traffic conditioning only at the ingress routers of the DiffServ domain. These ingress routers ensure that users do not violate the agreed-on traffic characteristics. Within the DiffServ domain, packets with the same DSCP are treated as the same class in each router, and all routers simply apply their local scheduling policies to the packets based on their DSCPs. A packet forwarding behaviour corresponding to a DSCP is called a per-hop behaviour (PHB). Each PHB is implemented in each router *locally* by means of buffer management and packet scheduling mechanisms, and the same DSCP may be mapped into different PHBs in different routers. Only two PHBs have been standardised: *expedited forwarding* [27] and *assured forwarding* [28]. The expedited forwarding requires a guaranteed amount of bandwidth for providing a low delay, low jitter and low loss service. Therefore it is appropriate for real-time services such as voice and video streaming. The assured forwarding is actually a group of PHBs. It has four classes and within each there are three levels of drop precedence. The assured forwarding focuses on the reliability of packet delivery, whilst delay and jitter are not as important as packet loss. Therefore, it is appropriate for non-real time services such as file transfer.

Although the DiffServ architecture provides a scalable QoS throughout the network, it fails to be the solution for end-to-end QoS provisioning because of traffic aggregation and different DSCP interpretations [29]. DiffServ also requires that the network administrators must use some mechanisms to install classification criteria in all the nodes in the DiffServ domain. Moreover, it is hard for DiffServ to anticipate traffic patterns and volumes so that QoS can be provided in real time [26].

To take advantage of the IntServ and DiffServ architectures, the IntServ over DiffServ architecture is proposed in [30]. This architecture uses DiffServ in the core network to avoid per-flow QoS signalling, and deploys IntServ in the access networks between the users and the core network so that users can explicitly express their QoS requirements. IntServ service requirements are mapped into DiffServ classes at the boundary between the access and core networks, and the DiffServ domain is treated like a single logical link between two IntServ-compliant networks.

2.3.3 Inefficiencies of QoS Architectures in Mobile Wireless Networks

As the IntServ and DiffServ architectures were initially designed for wired networks, they become inefficient when they are deployed in mobile wireless networks. For IntServ, this inefficiency is caused by its dependence on using RSVP signalling

for per-flow resource reservations [31]. When a mobile node changes its point of attachment to the network, a mobility management protocol such as MIPv6 will allocate it a new CoA. Since an RSVP reservation path is identified by the IP addresses of the communicating ends, the mobile node must request a new reservation path between itself and its correspondent node, and consequently its QoS session will be interrupted.

The inefficiency of DiffServ in mobile environments principally results from the service level agreement (SLA) negotiation and flow identification [32]. In DiffServ, the services a user receives are described and determined by a static SLA between the user and its Internet service provider. A mobile node at its home network carries out its QoS sessions depending on the SLAs that have been negotiated between its home network and its correspondent node's network. When the mobile node moves to another DiffServ domain, similar SLAs must be negotiated between the mobile node's visited network and its home network and/or between the visited network and the correspondent node's network, depending on how packets are delivered. Moreover, since packets may be classified using the IP addresses of the communicating ends, some mechanisms are required at the ingress routers of the DiffServ domain to determine whether the packets belong to a certain mobile node in order to mark them with the same DSCP and provide the same service to the mobile node as its previous network.

To provide QoS in mobile wireless networks, the integration of mobility and QoS management mechanisms is necessary, and this integration has been carried out for both IntServ and DiffServ. Compared to the DiffServ architecture, the IntServ architecture receives more interest because it provides end-to-end QoS and explicitly affects the management of network resources [26]. For the same reason, this thesis investigates the provisioning of integrated services in mobile environments, and more specifically, the performance modelling of the schemes that optimise RSVP-based network resource reservation. These schemes are proposed to solve the different problems associated with RSVP when it is used in mobile wireless networks, and they will be introduced in their respective chapters.

2.4 Handover Management in Heterogeneous Mobile Wireless Networks

Next-generation wireless communications are expected to require heterogeneity in terms of access technologies, services, capacities, etc. Taking account of the complementary characteristics of different wireless networks, especially in terms of bandwidth and coverage, the requirements of future wireless communications can be attained by the integration of heterogeneous wireless networks [33]. Since each individual network has its own characteristics, interworking different types of wireless networks requires careful design of network management schemes that provide seamless mobility, high QoS and strong security [34]. Moreover, users in this heterogeneous environment may not be satisfied with just reliable connectivity; they may also want to access their services through the best possible networks. The definition of *best* depends on a number of different aspects such as user preferences, device capacity, application demands and available network resources. To achieve this requirement, a handover management technique must include mechanisms that select appropriate networks for users [35].

Handover management is the process by which a mobile node keeps an active connection while moving from one point of attachment to another. In HMWNs, a mobile node may perform two types of handovers: *horizontal handovers* and *vertical handovers* [36]. A horizontal handover happens when the mobile node moves across cells that use the same type of access technology, whereas the movement between different types of wireless networks is referred to as a vertical handover. Both horizontal and vertical handover processes consist of two main phases: handover decision and handover execution [37, 38]. In the handover decision phase, all the information required to identify the need for a handover is collected and whether and how to perform the handover is determined. The handover execution phase largely consists of the allocation of resources at the mobile node's new point of attachment and re-routing its existing communication to the new location.

The major difference between the horizontal and vertical handover lies in the handover decision phase. When performing a horizontal handover, the mobile node generally makes its decision based on the evaluation of its received signal strength. For the vertical handover, however, there are various attachment options and the mobile node may choose a network according to its QoS requirements, service charges and

security associations, etc. Therefore, in addition to signal strength and availability that are used in horizontal handover, various handover criteria can be taken into account when making a vertical handover decision [38]:

- **Cost of services**: Cost is always a major consideration to users, and could sometimes be the decisive factor in a handover decision. Networks may employ different billing strategies that may influence the user's choice.
- **Network conditions**: Network-related parameters such as bandwidth, network latency and packet loss may be measured in order to make an effective network selection.
- **Mobile node conditions**: A mobile node's dynamic attributes such as mobility pattern, account balance and power consumption may be considered to facilitate proactive handover.
- **User preferences**: A user may have preference for one type of network over another.
- **Security**: The most significant source of risks in wireless networks is that the communication medium is open to intruders. Some types of services require strong security associations between the user and the service provider.

A possible problem of using multiple handover criteria is that the vertical handover decision may become difficult and ambiguous because each criterion may play a role in the decision making. To select a suitable network, it is necessary to evaluate each optional decision in terms of these criteria, and this evaluation process can be facilitated by formulating it as a mathematical expression, which is called a network selection strategy (NSS).

2.4.1 Network Selection Strategies

A number of NSSs have been proposed and they are generally based on multiple attribute decision making (MADM) theory. The widely used proposals are: simple additive weighting (SAW) based [39, 40], technique for order preference by similarity to ideal solution (TOPSIS) based [41] and fuzzy MADM based [42].

- **SAW-based**: In the SAW-based strategy, each network is associated with a point which is calculated as the weighted sum of all the handover related attribute

values. That is, for the candidate network i whose attribute values are x_{ij}, its point P_i is obtained by adding the *normalised* contributions from each attribute, ${r_{ij}}^{\text{v}}$, multiplied by its corresponding weight w_j:

$$P_i = \sum_{j=1}^{N} w_j * r_{ij},$$

where N is the number of handover related attributes and $\sum_{j=1}^{N} |w_j| = 1$. The weights for benefit attributes are positive and those for cost attributes are negative. All the candidate networks are then ranked according to their points and the network with the highest score is selected.

- **TOPSIS-based**: The TOPSIS-based strategy is based on the principle that the chosen network should have the shortest distance from the best solution and the longest distance from the worst solution. The best solution is the network whose attribute values x_{Bj} are the optimum among all the candidate networks. That is, $x_{Bj} = \max_{i \in 1,\cdots,M} x_{ij}$ or $x_{Bj} = \min_{i \in 1,\cdots,M} x_{ij}$, depending on whether x_{Bj} is a benefit attribute or a cost attribute. As for the worst solution, its attribute values x_{Wj} are chosen in the opposite way. The distances of the candidate network i from the best and the worst solutions, D_i^B and D_i^W, are then calculated as Euclidean distances:

$$D_i^B = \sqrt{\sum_{i=1}^{M} (x_{ij} - x_{Bj})^2}, \qquad D_i^W = \sqrt{\sum_{i=1}^{M} (x_{ij} - x_{Wj})^2}.$$

Based on the above distances, the preference for the network i, P_i, is defined as:

$$P_i = \frac{D_i^W}{D_i^B + D_i^W};$$

and the network with the highest preference is chosen.

- **Fuzzy MADM based**: In this type of strategy, fuzzy logic is used to represent the imprecise information of some network measures and user preferences. For example, the sojourn time of a mobile node in a cell can be described as short, medium and long. These linguistic terms can be converted to values ranging

[v] For example, r_{ij} can be calculated as $r_{ij} = \frac{x_{ij}}{\sum_{i=1}^{M} x_{ij}}$, where M is the number of candidate networks.

from 0 to 1 using a conversion function. After all the imprecise attributes are converted, ordinary MADM approaches can be used to select a network.

Among the multiple handover criteria that a vertical handover decision takes into account, some specific attributes are emphasised in some non-MADM based NSSs. In [43], user preference is of particular interest. Middleware sitting between the transport layer and the application layer is proposed to control the vertical handover behaviour. This middleware not only allows a user to change network but also enables an application to adapt to new network conditions. In [44], the proposed NSS first predicts the power consumption of a session using different network interfaces and then chooses the network which is the most power efficient. In [45], the only handover criterion is the achievable data rate, and the data rate of a network is represented by the received signal to interference and noise ratio (SINR). General reviews of the proposed strategies and their details can be found in [35,46,47] and the references therein.

In HMWNs, users can move between various networks with different characteristics. This heterogeneous environment provides the users with the potential for taking advantage of different access technologies according to their requirements. As the first phase of a vertical handover, the handover decision determines which network a mobile node will connect to and controls the session behaviour of the mobile node, and consequently, it affects both the performance of user applications at the mobile node end and the utilisation of network resources. Therefore, although NSSs generally operate at the application layer, they are still regarded as network management schemes. For that reason, in addition to schemes that provide integrated services in mobile wireless networks, this thesis also investigates performance modelling of different NSSs used in a particularly popular heterogeneous wireless network environment, i.e. 3G-WLAN interworking networks.

2.5 Summary

As a result of node mobility, network management in mobile wireless networks is more difficult than that in wired networks from the perspectives of mobility, QoS and heterogeneity. This chapter introduces the most widely used mobility management protocols used in mobile wireless networks, and the resultant inefficiency of the traditional QoS architectures. Moreover, the issue of vertical handover decisions in HMWNs, which is important to achieve ABC, is also reviewed. Of particular impor-

tance is the performance evaluation of the schemes that integrate mobility and QoS management in mobile environments and the NSSs designed for vertical handovers, because the evaluation can provide information on how much the proposed schemes improve the utilisation of network resources and the service quality perceived by users. To assess the performance of these network management schemes, performance models of them are required. In the next chapter, the modelling formalism used to construct these models will be introduced.

Chapter 3
Performance Modelling and Performance Evaluation Process Algebra

3.1 Introduction

The issue of performance modelling has concerned designers of computer, communication, and manufacturing systems throughout the history of their evolution. The increasing amount of functionality that is required to be performed by such systems has made their abstraction and evaluation an extremely complex task. This difficulty strongly depends on the capability of mapping the performance characteristics of the system components into the overall system-level performance characteristics [48]. To investigate whether a proposed design of a system performs as expected, models are developed. A *model* is an approximate and abstract representation of a system and hides its implementation details, and it allows analysis of some properties of the system before realising it in hardware and/or software. The process of constructing models and assessing performance properties of a system based on the models is referred to as the *performance modelling* of the system. Therefore, performance modelling is concerned with the capture and evaluation of the dynamic behaviour of systems, which involves the representation of system behaviour, and the definition and determination of characteristic performance measures.

This chapter introduces a performance modelling formalism, named Performance Evaluation Process Algebra (PEPA), which is employed in this thesis. Traditional performance modelling methods are reviewed in Section 3.2. In Section 3.3 a very short introduction to process algebra is presented as prerequisite background material for PEPA. Section 3.4 describes the syntax and semantics of PEPA, and demonstrates how PEPA can be used as a performance modelling technique through an example. Finally, Section 3.5 gives a summary of this chapter.

3.2 Methods for Performance Modelling

Except carrying out direct measurement on system performance through experiments, mathematical analysis techniques can also be applied to evaluating the performance of a system. Two such classical techniques are: simulations and mathematical modelling.

The simulation is a numerical technique for conducting sampling experiments of a system evolving in time [49]. A simulation model of a system describes the dynamic behaviour of the system in terms of individual events and interactions of the components of the system. Almost any level of detail about the system can be included in a simulation model, and no particular artificial assumption has to be imposed. Therefore, the simulation method is widely used for performance analysis, and especially for systems that are so complex that the mathematical modelling is not applicable. However, constructing, optimising, running, analysing and validation of simulation models are usually time-consuming; models including too many details are usually hard to understand and computationally expensive. In fact, simulation and mathematical modelling complement each other. Although it is hard to mix mathematical modelling and simulation, this combination can provide the efficiency of mathematical modelling and the realism of simulation modelling [48].

The mathematical modelling is a solution technique which extracts a functional relation between system parameters and a chosen performance criterion in terms of equations that are analytically solvable [48]. The most important mathematical modelling is queueing theory. Most performance problems, especially those in communication networks, are caused by contention for resources. The formation of a queue is a general consequence that occurs when requests for a service facility exceed the capacity of that facility. Queueing theory has been extensively studied and provides a mathematical framework for formulating and analysing the effect of various resource contentions [48, 50, 51]. Systems with one queue and one or more identical or different servers can be modelled as *single station* queueing systems and generally have straightforward and computationally efficient solutions. Complicated systems with multiple resources can be modelled as *queueing networks*, which consist of networks of single station queueing systems, using a hierarchical modelling approach which decomposes the system into micro-level models in a stepwise fashion. However, only the complex systems that have certain structures

and follow some assumptions have *product form* solutions, and the derivation of performance measures is not easy and the results are usually approximations. In brief, although queueing theory is very powerful when applicable, it has limited expressiveness for systems mainly because of some underlying assumptions [51]. Furthermore, modelling systems as queueing models is an art mastered only by experienced specialists, and complex queueing models are often not mathematically tractable even after simplification and decomposition [52, 53].

Even though queueing theory is elaborate and have many fundamental results, the size and complexity of many modern systems expose the deficiencies of it in expressing concurrency and interdependency [52]. This leads to recent interest in approaches that can tackle the above problems and consider both functional and temporal aspects of a system at the same time. One such approach is stochastic Petri net (SPN) which is an effective modelling formalism for describing and analysing the flow of workload and control in a system [54]. Two elementary nodes of SPN are *places* and *transitions*, and a finite number of *tokens* can be distributed among places. In SPN, the states and behaviour of a system are expressed as the distribution of tokens in places and the transitions between places, which helps in understanding and reasoning about the functional aspects of the system. By using exponential distributions for the temporal specifications of transitions, i.e. exponentially timed transitions, an underlying Markov chain is associated with the model from which performance measures can be derived. To extend the expression capability of SPNs, *immediate* transitions are introduced in SPNs representing probabilistic choices that do not consume time, and this type of SPN is known by the name generalised stochastic Petri nets (GSPNs) [55]. There are also SPNs using general distributions but the analytical or numerical analysis of them is impossible and simulation is the only way to analyse them [56]. The major limitations of SPNs and GSPNs are that they are expensive in specifying and representing complex systems, and it is easy to construct large models that are difficult to understand and numerically intractable. Moreover, they lack the *compositionality* property, which is very important, enabling the construction of complex performance models from small building blocks [57].

The compositionality property is captured by stochastic process algebras (SPAs) which have been developed to describe and evaluate resource-sharing systems. In an SPA, a model is compositionally constructed from submodels that exhibit specific

behaviour. The functional and temporal aspects of a system are then formulated by the interactions and interdependency between these submodels. Moreover, an SPA uses a high-level language to describe systems, which facilitates model construction. It has formal *semantics* that can translate a system description into a state transition system which is associated with an underlying stochastic process. Performance evaluation of the system is then carried out based on this stochastic process whose characteristics are dependent on the types of random distributions that are used in the system description. As in SPNs, using general distributions enriches the expressiveness of an SPA but then it is not supported by general analysis techniques. For that reason, the exponential distribution is usually used, which makes the underlying process a Markov chain and simplifies performance analysis. The explicit compositionality property makes SPAs stand out from existing performance modelling paradigms since it facilitates constructing complex performance models. Components and their interactions in a system can be modelled separately; the structure of a model is easily constructed and clear to understand; components of a model can be transplanted from another model without making the model intractable. Many case studies demonstrating the benefits of SPAs can be found in the references in [58, 59]. In the light of the above benefits, a particular SPA, *Performance Evaluation Process Algebra* (*PEPA*), has been chosen as the modelling technique used in this thesis. In the following sections, introductions to process algebras and PEPA are provided.

3.3 Process Algebras

Process algebras have been designed as formal description techniques for concurrent systems — systems that consist of subsystems interacting with each other. A process algebra theory introduces processes as the basic terms of an algebraic language which comprises a small number of combinators. In process algebras, a *process* or an *agent* can perform *actions*, and a system is modelled from the perspective of its *behaviour* as interactions between processes. By using some basic axioms similar to those in elementary algebra[1], equational reasoning can be carried out in order to decide whether two systems are behaviourally equivalent, or to verify that a system satisfies some properties, or to investigate other aspects of a system. In short, process algebra

[1]The simplest examples are $a + b = b + a$, $a + (b + c) = (a + b) + c$, etc. Here, the combinator "+" denotes the *choice* operation rather than the *addition* operation, which will be explained later.

is the study of the behaviour of systems by algebraic means [60].

The model of a system, in classical process algebras such as Communicating Sequential Processes (CSP) [61] and Calculus of Communicating Systems (CCS) [62], is composed of processes which model the component parts of the system. *Combinators* of the language are used to describe the behaviour of processes. For example, the *prefix* combinator "." designates a first action of a process, e.g. the process $\alpha.P$ evolves into process P after performing an α action. In classical process algebras a system can be expressed as a group of processes which undertake actions independently or interact with each other. The operational semantics of the language makes it possible to construct a labelled transition system for a model, in which processes form the nodes and actions form the arcs. This structure is a useful tool for reasoning about the behavioural properties of a system. However, since the objective of classical process algebras is qualitative analysis rather than quantitative, an action is instantaneous and time information is abstracted away. This limitation means that quantitative values such as response time and throughput cannot be extracted from classical process algebra models.

To add quantitative information like time and probability to classical process algebras so that they are suitable for performance modelling, *timed* and *probabilistic* process algebras have been developed. The main idea of timed process algebras is to associate time information with an action, and that of probabilistic process algebras is to calibrate the uncertainty of a process' behaviour with a probabilistic choice operator. Timed and probabilistic process algebras can be regarded as logical predecessors of SPAs, in which time and probability are integrated together into process algebras by associating a probability distribution function with each action representing its time information. Typical examples of SPAs are PEPA [63], TImed Process for Performance Evaluation (TIPP) [64] and Extended Markovian Process Algebra (EMPA) [65]. PEPA was the first process algebra to be developed with the intention of generating Markov chains which could be solved numerically for performance evaluation. It is used as the performance modelling technique in this thesis.

3.4 Performance Evaluation Process Algebra

This section gives an introduction to the PEPA language, starting with the basic terms and execution rules of the language. How a stochastic process representation of

a system is generated from its PEPA model is explained. Moreover, an example is used to illustrate how the language can describe a system and how performance measures can be derived from a PEPA model.

3.4.1 Components and Activities

PEPA is an SPA which provides a compositional approach to performance modelling. In PEPA a system is described as an interaction of *components* which engage in *activities*. A PEPA model has a countable number of components that correspond to identifiable parts of a system. For example, a printing system can be considered to consist of a *Queue* component which buffers jobs and a *Printer* component which prints jobs. This abstract description of a system is similar to the design of a system and facilitates model construction.

The behaviour of a component is captured by its activities. The *Queue* component in the above example could have activities *arrive* and *print*, representing a job entering and leaving the queue respectively. In PEPA, an activity of a component is characterised by its type and duration. For example, an activity a is defined as a pair (α, r), where α is the *action type* (or simply *type*) and r is the *activity rate* (or simply *rate*). Each activity has only one type, and its rate is the parameter of an exponential distribution governing its duration.[II,III] For a component P, the set of all its enabled action types and the set of all its enabled activities are denoted as $\mathcal{A}(P)$ and $\mathcal{A}ct(P)$ respectively. Since multiple enabled activities of P can have the same action type, $\mathcal{A}(P)$ is a set whereas $\mathcal{A}ct(P)$ is a *multiset*. If the component P behaves as component P' after completing activity $a \in \mathcal{A}ct(P)$, the P' is called a *derivative* of P and this transition can be denoted as $P \xrightarrow{a} P'$, or $P \xrightarrow{(\alpha,r)} P'$.

3.4.2 Syntax and Execution Rule

The PEPA language provides a small set of combinators, which are used to express the individual behaviour of components and the interactions between them. The brief definitions and interpretations of these combinators are given below.

Prefix: $(\alpha, r).P$

The prefix combinator "." designates activities of a component. The component $(\alpha, r).P$ carries out activity (α, r) and subsequently behaves as component P. Sequen-

[II]The rate r can be any positive real number, and the duration t of activity a follows the distribution function $F_a(t) = 1 - e^{-rt}$.

[III]The justification of the exponential distribution assumption will be discussed in Section 3.4.7.

tial activities of a component can be expressed by concatenating them with the prefix combinator. For example, the *Printer* component can print a job and then suspend for a while before it is ready to take the next job. This behaviour can be expressed as:

$$Printer \stackrel{def}{=} (print, r_1).(suspend, r_2).Printer$$

or equivalently as:

$$\begin{aligned} Printer &\stackrel{def}{=} (print, r_1).Printer' \\ Printer' &\stackrel{def}{=} (suspend, r_2).Printer \end{aligned}$$

where the component $Printer'$ denotes the behaviour of the *Printer* component after completing the *print* activity.

Choice: $P + Q$

The choice combinator "+" expresses uncertainty about the behaviour of a component. All the enabled activities in components P and Q are enabled in the component $P + Q$ and they compete with each other for completion. The first activity to be completed must be an activity of P or Q, and this activity distinguishes one of the components, P or Q, and $P + Q$ will subsequently behave as this component. The continuous nature of the probability distributions ensures that P and Q cannot complete an activity at the same time. For example, let the component $Queue_i$ denote the behaviour of the *Queue* component when there are i jobs in the queue. It can either allow another job to arrive (when the queue is not full) or have one of its jobs printed (when the queue is not empty). The *Queue* component can then be defined as $(i = 1, 2, \cdots, N-1)$:

$$\begin{aligned} Queue_0 &\stackrel{def}{=} (arrive, r_3).Queue_1 \\ Queue_i &\stackrel{def}{=} (arrive, r_3).Queue_{i+1} \\ &\quad + (print, \top).Queue_{i-1} \\ Queue_N &\stackrel{def}{=} (print, \top).Queue_{N-1} \end{aligned}$$

where N is the maximum size of the queue. The symbol "$\top$" means the rate of the activity is outside the control of the component. In this example, the *Queue* component is *passive* with respect to the activity *print* since it cannot influence the rate at which jobs are printed.

Hiding: P/L

The hiding combinator "/" is used to abstract activities of a component. The component P/L behaves as P except that any activities of types within set L are not identifiable externally. A hidden activity appears as the *unknown* type τ but its duration is unaffected.[IV] For example, the activity *suspend* may be hidden from the outside, and this can be achieved by redefining the *Printer'* component as:

$$Printer' \stackrel{def}{=} (suspend, r_2).Printer/\{suspend\}$$

which is equivalent to:

$$Printer' \stackrel{def}{=} (\tau, r_2).Printer$$

Cooperation: $P \underset{L}{\bowtie} Q$

The cooperation combinator " $\underset{L}{\bowtie}$ " represents the interaction between components P and Q. The set L is called the *cooperation set* and it defines a set of action types that must be carried out by P and Q *together*. The activities of P and Q whose action types are not included in set L are called *individual* activities and they will proceed unaffected. In contrast, all activities whose action types occur in set L are called *shared* activities and will only proceed when they are enabled in both P and Q. Therefore, it is possible that one component becomes blocked and has to wait for its partner to be ready to participate in cooperation. When the cooperation set L is empty, the two components proceed concurrently without any interaction between them. A shorthand notation $P \| Q$ is used to represent $P \underset{\varnothing}{\bowtie} Q$, and the symbol "$\|$" is referred to as the *parallel* combinator. For example, the printing system may have two parallel queues that share only one printer, and each *Queue* component cooperates with the *Printer* component on the activity *print* individually. This can be expressed as:

$$(Queue_0 \| Queue_0) \underset{\{print\}}{\bowtie} Printer$$

The rate of a shared activity is determined by the rate of the slower participant and is thus the smaller of the two rates. If one of the participants is passive with respect to the cooperation, i.e., its activity rate is labelled as $\top$, the rate of the shared activity is determined by the other component. This means that although the component which

[IV] The action type τ is reserved in the PEPA language to represent activities that are unknown or unimportant.

has this passive activity is required to engage in the cooperation, it has no influence on the rate at all.

Constant: $A \stackrel{def}{=} P$

The constant combinator "$\stackrel{def}{=}$" can be used to associate names with behaviour. Here A is the constant and it is given the behaviour of the component P. Its usage has been shown in the above examples. Moreover, it can be used for *system definition* which specifies how the system is constructed from the defined components, i.e., how the components cooperate with each other so that they express the behaviour of the system. For example, the printing system with one queue and one printer can be given a name *System*, which is associated with the cooperation between the components *Queue* and *Printer*. That is:

$$System \stackrel{def}{=} Queue_0 \underset{\{print\}}{\bowtie} Printer$$

where $Queue_0$ and *Printer* define the *initial behaviour* of the corresponding components.

Precedence:

The precedence of the above combinators is as follows: hiding has the highest precedence followed by prefix, cooperation comes next and choice has the lowest precedence. Brackets may be used to force a different precedence or to clarify the grouping as in elementary algebra.

Execution Rule:

In PEPA the dynamic behaviour of a component when it has more than one activity enabled is governed by a rule called *race condition*. That is, all the enabled activities compete with each other to proceed but only the *fastest* succeeds. The probability that an activity wins the race is given by the ratio of the rate of this activity to the sum of the activity rates of all the enabled activities [63]. Therefore, the race condition is the mechanism by which PEPA achieves probabilistic behaviour. This property is useful when an action being modelled has more than one outcome. For example, a component P which is defined as:

$$P \stackrel{def}{=} (\alpha, \tfrac{1}{3} * r).P_1 + (\alpha, \tfrac{2}{3} * r).P_2$$

has two separate activities of the same action type. To external observers, the

component P engages in activities of type α with rate r, and will subsequently behave as P_1 with probability $1/3$ or as P_2 with probability $2/3$. The sum of the rates of all activities of an action type, say α, in component P is called the *apparent rate* of action of type α, and is denoted as $r_\alpha(P)$.

3.4.3 Operational Semantics

Prefix

$$\frac{}{(\alpha, r).P \xrightarrow{(\alpha,r)} P}$$

Choice

$$\frac{P \xrightarrow{(\alpha,r)} P'}{P + Q \xrightarrow{(\alpha,r)} P'} \qquad \frac{Q \xrightarrow{(\alpha,r)} Q'}{P + Q \xrightarrow{(\alpha,r)} Q'}$$

Hiding

$$\frac{P \xrightarrow{(\alpha,r)} P'}{P/L \xrightarrow{(\alpha,r)} P'/L} \quad (\alpha \notin L) \qquad \frac{P \xrightarrow{(\alpha,r)} P'}{P/L \xrightarrow{(\tau,r)} P'/L} \quad (\alpha \in L)$$

Cooperation

$$\frac{P \xrightarrow{(\alpha,r)} P'}{P \underset{L}{\bowtie} Q \xrightarrow{(\alpha,r)} P' \underset{L}{\bowtie} Q} \quad (\alpha \notin L) \qquad \frac{Q \xrightarrow{(\alpha,r)} Q'}{P \underset{L}{\bowtie} Q \xrightarrow{(\alpha,r)} P \underset{L}{\bowtie} Q'} \quad (\alpha \notin L)$$

$$\frac{P \xrightarrow{(\alpha,r_1)} P' \quad Q \xrightarrow{(\alpha,r_2)} Q'}{P \underset{L}{\bowtie} Q \xrightarrow{(\alpha,R)} P' \underset{L}{\bowtie} Q'} \quad (\alpha \in L) \quad \text{where } R = \frac{r_1}{r_\alpha(P)} * \frac{r_2}{r_\alpha(Q)} * \min(r_\alpha(P), r_\alpha(Q))$$

Constant

$$\frac{P \xrightarrow{(\alpha,r)} P'}{A \xrightarrow{(\alpha,r)} P'} \quad (A \stackrel{def}{=} P)$$

Figure 3.1: *Operational Semantics of PEPA*

The semantics of PEPA can be presented in the structural operational semantics (SOS) [66] style as shown in Figure 3.1. Operational semantics of a process algebra provide a formal interpretation of all expressions, and explicitly describe how processes evolve in stepwise fashion and possible state transitions they perform. Transitions are governed by operational rules which have the form $\frac{\text{premises}}{\text{conclusions}}$, and the operational rules in Figure 3.1 can be read as: if the transition(s) above the inference line can be inferred, then the transition below the line can be deduced.

In the rule for cooperation, the apparent rate of a shared activity in the component

$P \underset{L}{\bowtie} Q$, i.e. $r_\alpha(P \underset{L}{\bowtie} Q)$, is taken to be the smaller of the apparent rates of that action type in the components P and Q, i.e. $\min(r_\alpha(P), r_\alpha(Q))$. Moreover, P and Q may have multiple activities of the same action type which have different outcomes. In this case, the activity (α, R) is in fact an instance of the action type α of the component $P \underset{L}{\bowtie} Q$. The probability that the activity (α, r_1) in P and the activity (α, r_2) in Q are combined to form the shared activity is $r_1/r_\alpha(P) * r_2/r_\alpha(Q)$.[V] It is clear that, if P and Q have only one instance of the action type α, then R is $\min(r_1, r_2)$.

On the basis of the above semantic rules, PEPA can be defined as a *labelled multi-transition system*. In general, a labelled transition system $(S, T, \{\xrightarrow{t} | t \in T\})$ consists of a set of states S, a set of transition labels T and a transition relation between states $\xrightarrow{t}$ for each $t \in T$. Since PEPA allows multiple instances of a transition, i.e. multiple activities of the same action type, it can be regarded as a labelled multi-transition system $(\mathcal{C}, \mathcal{A}ct, \{\xrightarrow{(\alpha,r)} | (\alpha, r) \in \mathcal{A}ct\})$, where $\mathcal{C}$ is a set of components, $\mathcal{A}ct$ is a set of activities and $\xrightarrow{(\alpha,r)}$ is a set of transitions governed by operational rules in Figure 3.1.

Following these rules, a *transition diagram* of a PEPA model which shows its behaviour can be constructed. Figure 3.2 shows the transition diagram of the above printing system example with one queue and one printer when the maximum queue length is 3.

3.4.4 The Continuous-time Markov Chains Underlying PEPA Models

For any PEPA model, an underlying stochastic process can be generated based on its transition diagram. To form the stochastic process, a state is associated with each node of the diagram (e.g. the component $Queue_2 \underset{\{print\}}{\bowtie} Printer$), and the transitions between states are defined by the arcs of the diagram (e.g. the activity $(print, r_1)$). Since all activities are time homogeneous and their durations are exponentially distributed, it is established that this stochastic process has the *Markov property* and is a continuous-time Markov chain (CTMC) [63]. The elements of the infinitesimal generator matrix of the underlying CTMC, $\mathbf{Q}$, can be defined intuitively as follows:

- The off-diagonal transition rate of $\mathbf{Q}$ between the two states corresponding to the components C_i and C_j $(i \neq j)$, q_{ij}, is defined as $q(C_i, C_j) = \sum_{a \in \mathcal{A}ct(C_i|C_j)} r_a$, where $\mathcal{A}ct(C_i|C_j)$ is the multiset of activities connecting the components C_i and

[V] Choices in P and Q are assumed to be independent.

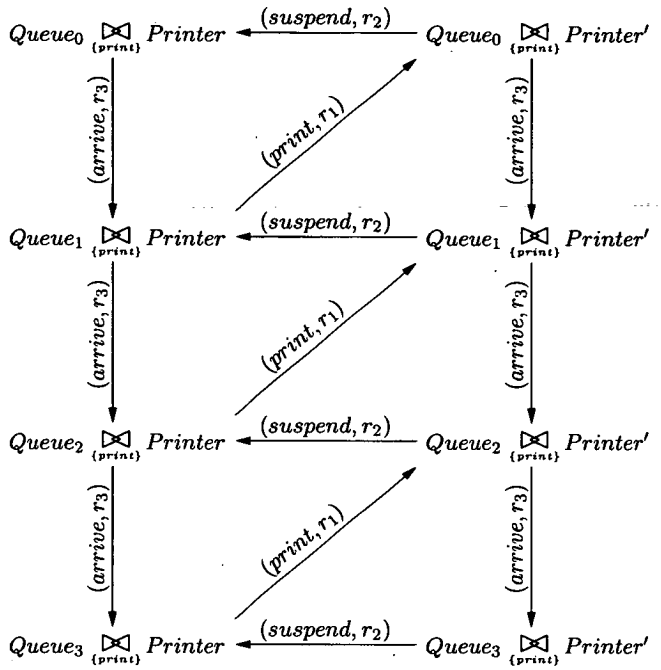


Figure 3.2: *The transition diagram of the PEPA model of the printing system*

C_j. This can be regarded as the rate at which the system changes from behaving as C_i to behaving as C_j.

- The sojourn time of the CTMC in the state corresponding to the component C_i is exponentially distributed, and the rate at which the system leaves this state is defined as $q(C_i) = \sum_{a \in Act(C_i)} r_a$, where $Act(C_i)$ is the multiset of activities leaving the components C_i. The diagonal element of $\mathbf{Q}$ corresponding to the component C_i is the negative value of $q(C_i)$, i.e. $q_{ii} = -q(C_i)$.

N.B.: Without ambiguity, throughout the rest of this thesis, a PEPA model behaving as a component C is said to be in the *state* C unless otherwise stated.

To ensure the CTMC underlying a PEPA model has the *steady state* or *equilibrium* behaviour, the PEPA model must be *cyclic* and *finite* [63]. From the perspective of the transition diagram of a PEPA model, these two requirements mean that any node of the diagram can be reached from any other node of the diagram, and that the diagram has finite nodes, respectively. The first requirement is satisfied if all components of a PEPA model are cyclic; its system definition only consists of cooperation between these components; and the cooperation sets in its system definition are well defined so that there is no deadlock in the model's transition diagram. The second requirement

is satisfied if the PEPA model consists of finite number of components.[VI]

3.4.5 Deriving Performance Measures

In this thesis, only steady state performance analysis is considered. Performance measures such as utilisation and throughput of a PEPA model can be derived from the *equilibrium probability vector* of its underlying CTMC, i.e. the equilibrium probabilities that the PEPA model is in each state. Assume the state space of the PEPA model is S and π denotes the equilibrium probability vector of the model, then π can be found by using the *global balance equations* (Eq. (3.1)) and the *normalisation condition* (Eq. (3.2)):

$$\mathbf{Q}\boldsymbol{\pi} = \mathbf{0}, \tag{3.1}$$

$$\sum_{C_i \in S} \pi(C_i) = 1, \tag{3.2}$$

where $\pi(C_i)$ is the equilibrium probability of the model in the state C_i. Direct methods such as Gaussian elimination and iterative methods such as Gauss-Seidel method can be used to solve this *system of linear equations* [67].[VII] $\pi(C_i)$ can be regarded as the probability that the system is in the state C_i when it is observed at a random time. Alternatively, it can be considered as the proportion of time that the system behaves as the component C_i.

Performance measures can be derived using a Markov reward structure (MRS) [68], which assigns rewards to the states or transitions of interest of the Markov chain. For PEPA models, rewards can be assigned to *activities* of a component. Then the reward associated with the state corresponding to this component is the sum of the rewards attached to the activities that the component enables. A performance measure is then calculated as the total reward R using Eq. (3.3):

$$R = \sum_{C_i \in S} \rho_i * \pi(C_i), \tag{3.3}$$

where ρ_i is the reward associated with the state C_i. The term $\rho_i * \pi(C_i)$ can be regarded as the reward the system accrues during its sojourn time in state C_i. In this way, the rewards can be defined at the level of the PEPA model, rather than at the level of the

[VI] All the PEPA models in this thesis have equilibrium behaviour.

[VII] For all the PEPA models in this thesis, their equilibrium probability vectors are calculated using a solver that implements the Gauss-Seidel method.

underlying Markov chain.

PEPA is equipped with tools for manipulating and analysing models, and thus analysis of a PEPA model becomes automatic once its description is completed. All the evaluation in this thesis has been conducted using the PEPA Workbench and associated tools [69], and a review of these tools can be found in [59].

3.4.6 An Example

In this subsection, the process of solving a PEPA model and deriving performance measures from it are demonstrated, using the printing system example introduced earlier. The PEPA model of the printing system is

$$\begin{aligned} Printer &\stackrel{def}{=} (print, r_1).Printer' \\ Printer' &\stackrel{def}{=} (suspend, r_2).Printer \end{aligned} \qquad \begin{aligned} Queue_0 &\stackrel{def}{=} (arrive, r_3).Queue_1 \\ Queue_i &\stackrel{def}{=} (arrive, r_3).Queue_{i+1} \\ &\quad + \ (print, \top).Queue_{i-1} \\ Queue_N &\stackrel{def}{=} (print, \top).Queue_{N-1} \end{aligned}$$

$$System \stackrel{def}{=} Queue_0 \underset{\{print\}}{\bowtie} Printer$$

If the maximum size of the queue N is 3, then this model has 8 states and 13 transitions, as shown in Figure 3.2. The 8 states are labelled as S_1, S_2, $\cdots$, S_8, which are identified as follows:

$$\begin{aligned} S_1 &\rightarrow Queue_0 \underset{\{print\}}{\bowtie} Printer \\ S_2 &\rightarrow Queue_1 \underset{\{print\}}{\bowtie} Printer \\ S_3 &\rightarrow Queue_2 \underset{\{print\}}{\bowtie} Printer \\ S_4 &\rightarrow Queue_3 \underset{\{print\}}{\bowtie} Printer \end{aligned} \qquad \begin{aligned} S_5 &\rightarrow Queue_0 \underset{\{print\}}{\bowtie} Printer' \\ S_6 &\rightarrow Queue_1 \underset{\{print\}}{\bowtie} Printer' \\ S_7 &\rightarrow Queue_2 \underset{\{print\}}{\bowtie} Printer' \\ S_8 &\rightarrow Queue_3 \underset{\{print\}}{\bowtie} Printer' \end{aligned}$$

The generator matrix **Q** of the model is

$$
\begin{array}{c}
 \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \\ S_7 \\ S_8
\end{array}
\begin{array}{c}
\begin{array}{cccccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 & S_7 & S_8 \end{array} \\
\left[\begin{array}{cccccccc}
-r_3 & r_3 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -(r_1+r_3) & r_3 & 0 & r_1 & 0 & 0 & 0 \\
0 & 0 & -(r_1+r_3) & r_3 & 0 & r_1 & 0 & 0 \\
0 & 0 & 0 & -r_1 & 0 & 0 & r_1 & 0 \\
r_2 & 0 & 0 & 0 & -(r_2+r_3) & r_3 & 0 & 0 \\
0 & r_2 & 0 & 0 & 0 & -(r_2+r_3) & r_3 & 0 \\
0 & 0 & r_2 & 0 & 0 & 0 & -(r_2+r_3) & r_3 \\
0 & 0 & 0 & r_2 & 0 & 0 & 0 & -r_2
\end{array}\right]
\end{array}.
$$

The activity rates used in the model are $r_1 = 1/10$, $r_2 = 1/3$ and $r_3 = 1/15$. By solving the system of equations Eq. 3.1 and Eq. 3.2, the equilibrium probability vector of the model is

$$
\begin{array}{c}
 \\ \boldsymbol{\pi} =
\end{array}
\begin{array}{c}
\begin{array}{cccccccc} \pi(S_1) & \pi(S_2) & \pi(S_3) & \pi(S_4) & \pi(S_5) & \pi(S_6) & \pi(S_7) & \pi(S_8) \end{array} \\
\left[\begin{array}{cccccccc}
0.2751 & 0.2201 & 0.1834 & 0.0550 & 0.1541 & 0.0550 & 0.0477 & 0.0096
\end{array}\right]
\end{array}.
$$

As examples, utilisation and throughput of the printer, and the average queue length are derived from this model.

- **Utilisation**: The utilisation of the printer is the proportion of time the printer has jobs to print. This can be calculated by assigning a reward of 1 to each state in which the component *Printer* is able to print jobs [70]. Therefore, the reward vector $\boldsymbol{\rho}$ is

$$
\begin{array}{c}
 \\ \boldsymbol{\rho} =
\end{array}
\begin{array}{c}
\begin{array}{cccccccc} \rho_1 & \rho_2 & \rho_3 & \rho_4 & \rho_5 & \rho_6 & \rho_7 & \rho_8 \end{array} \\
\left[\begin{array}{cccccccc}
0 & 1 & 1 & 1 & 0 & 0 & 0 & 0
\end{array}\right]
\end{array},
$$

  and the utilisation of the printer U is calculated as:

$$
U = \sum_{i=1}^{8} \rho_i * \pi(S_i) = 0.4585 = 45.85\%.
$$

- **Throughput**: The throughput of the printer is the expected number of printed jobs per unit time, i.e. the throughput of the activity *print*. Accordingly, it is an activity related performance measure and a reward equal to the activity rate r_1 is associated with each enabled activity *print* [71]. That is, the rewards assigned to each state are

$$\rho = \begin{matrix} \rho_1 & \rho_2 & \rho_3 & \rho_4 & \rho_5 & \rho_6 & \rho_7 & \rho_8 \end{matrix} \\ \begin{bmatrix} 0 & 1/10 & 1/10 & 1/10 & 0 & 0 & 0 & 0 \end{bmatrix},$$

and the throughput of the printer T is calculated as:

$$T = \sum_{i=1}^{8} \rho_i * \pi(S_i) = 0.0459.$$

- **Average Queue Length**: The average queue length is the expected number of queued jobs. This performance measure is the weighted sum of the equilibrium probabilities of system states, where the weights are the number of queued jobs in the corresponding states [72]. Consequently, the rewards assigned to each state are

$$\rho = \begin{matrix} \rho_1 & \rho_2 & \rho_3 & \rho_4 & \rho_5 & \rho_6 & \rho_7 & \rho_8 \end{matrix} \\ \begin{bmatrix} 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 \end{bmatrix},$$

and the queue length L is calculated as:

$$L = \sum_{i=1}^{8} \rho_i * \pi(S_i) = 0.9311.$$

3.4.7 PEPA for Performance Modelling

As presented in Section 3.2, the most important motivation for the use of process algebras for performance modelling is the compositionality property. The process algebraic combinators of PEPA are carefully designed so that they can express the compositional structures of systems. This enables the construction of complex systems as the combination of conceptually simpler subsystems, which is close to the way that designers think about systems. Moreover, PEPA is a high-level language for describing systems and thus is much more efficient than directly developing a state transition diagram of a system that is error-prone and time-consuming.

PEPA models are generally solved numerically at the level of the underlying Markov chains. The exponential distribution used in PEPA models reduces the mathematical complexity and makes it straightforward to derive a CTMC from a given specification. Although the exponential distribution assumption is usually not met in practice, this simplified assumption is still used as a building block of this thesis. On

the other hand, studies in performance modelling of stochastic systems have shown that results derived from this unrealistic assumption are often found to agree well with field data and experiments. That is, it is justified from an engineering point of view for performance evaluation of stochastic systems [73]. Moreover, stochastic systems may be insensitive or robust with assumptions made on system parameters [74].

Performance modelling techniques which rely on numerical solution are generally prone to the *state space explosion* problem — the size of the underlying Markov chain is so large that the model is intractable. However, applying equivalence reasoning on PEPA models allows model simplification and aggregation which results in reduction in the number of states required in the underlying Markov chain to represent the model. This kind of manipulation on models is also a distinguishing property of process algebras which alleviates the state space explosion problem.[VIII]

3.5 Summary

Performance modelling has been playing an important role in the design and evaluation of systems such as computers and communication networks. Efficient and reliable design and evaluation of complex systems like these require formal modelling approaches. The aim of this chapter is to give a brief review of the commonly used performance modelling methods, and introduce the PEPA formalism which is used in this thesis. Essentially, mobile wireless networks are resource-sharing systems which entail concurrency (multiple users compete for, and interact with, multiple resources), and compositionality (users, protocols and network entities are component parts of networks). The compositionality, concurrency and parsimony of PEPA make it suitable for performance modelling of the network management schemes that are used in mobile wireless networks. Models of these schemes are constructed and evaluated in the following chapters.

[VIII] All the PEPA models in this thesis are compiled by the PEPA tools that implement this state space reduction function.

Chapter 4
Modelling of Signalling Schemes for RSVP in Mobile Wireless Networks

As discussed in Chapter 2, RSVP exhibits deficiencies when it is deployed in mobile wireless networks. One of the major problems is the signalling optimisation problem with regard to reducing handover interruptions and overheads. In this chapter, PEPA models of basic and optimised signalling schemes for RSVP in mobile wireless networks are built and evaluated.

4.1 Introduction

In a mobile scenario, RSVP becomes inefficient because there is an interruption of a mobile node's QoS session when it changes its point of attachment to the network. A number of variants of the basic RSVP with mobility support have been proposed to tackle incompatibilities between mobility and QoS management protocols. Although most of the mobility-supported RSVP proposals integrate RSVP with the macro-mobility management protocols such as MIPv6, a mobility-supported RSVP that integrates RSVP and micro-mobility management protocols has become widely accepted as the best approach to combining mobility and QoS. This is because a micro-mobility management protocol such as HMIPv6 has inherent characteristics which facilitate the deployment of RSVP in mobile wireless networks. However a mobility-supported RSVP is designed, its essential objective is to optimise the signalling procedure of the basic RSVP by restricting its signalling to within the affected part of the network, so that the interruptions during handover can be reduced.

To identify the advantages of the mobility-supported RSVP, most of the previous efforts to evaluate the enhancements have been simulation in which specific network topologies and traffic patterns are used, and performance measures such as handover signalling delay and packet loss are obtained. The contribution of the work presented

in this chapter is that the PEPA models are the first formal performance models of the operation of both the basic and mobility-supported RSVP in mobile environments. Moreover, these models are independent of the specific implementations of the signalling schemes and the structure and behaviour of these PEPA models exhibit clear representations of the mechanisms underlying the schemes. From the PEPA models, important performance measures such as handover blocking probability and handover signalling cost are derived from the stationary behaviour of a mobile node, and the benefits of the mobility-supported RSVP are demonstrated.

The rest of this chapter is structured as follows. Section 4.2 gives a general review of the approaches that are used by mobility-supported RSVP to optimise the signalling procedure of the basic RSVP in mobile wireless networks. In Section 4.3 the PEPA models of both basic and mobility-supported RSVP are presented. The performance of the two schemes are evaluated and compared in Section 4.4 and in Section 4.5 the evaluation results are discussed.

4.2 Signalling Optimisation Schemes for RSVP in Mobile Wireless Networks

Since an RSVP reservation path is identified by the source and destination addresses, one of the major incompatibilities between RSVP and mobility management, when providing QoS guarantees in mobile wireless networks, is that a mobile node must re-establish an RSVP reservation path whenever it performs a network layer handover. This procedure causes interruptions to the mobile node's ongoing communications and can significantly degrade QoS-sensitive services. However, in most realistic situations, the old and new reservation paths between the mobile node and its correspondent node before and after a handover share some intermediate routers. Therefore, to reduce the reservation re-establishment time, it is required that the RSVP signalling should be localised within the affected part of the network [75]. Since the work of this chapter is not on designing a signalling optimisation scheme, only a general description of the approaches to localising the RSVP signalling during handovers is presented. The details of the approaches can be found in their corresponding references [76–86].

A straightforward way to localise the RSVP signalling is to configure an RSVP-compliant router so that it is able to determine whether it has already reserved

resources for an RSVP session. In mobility-supported RSVP schemes for MIPv6 [76–79], the flow label [87] in the IPv6 packet header is used to identify a traffic flow. When a mobile node performs a handover, it sends a reservation request which includes the flow label. The request is transmitted upstream to the mobile node's correspondent node, and along the path the RSVP-compliant routers will examine the flow label. If a router finds that there has been a reservation path for the flow, the reservation request is ignored.

Although the above mentioned work studies the integration of RSVP with MIPv6, it is indicated in [88] that for every mobile node's network layer handover, a micro-mobility management protocol such as HMIPv6 is preferable to a global-mobility management protocol such as MIPv6 from the perspectives of update latency, signalling overheads and location privacy. Therefore, schemes that integrate RSVP and HMIPv6 receive more interest. Previous work on deploying RSVP in an HMIPv6-compliant network [80–82] takes advantage of the two-layer care-of addresses of a mobile node. As presented in Chapter 2, when the mobile node performs a micro handover, i.e., switches to a new AR connecting to the same MAP, only its LCoA is changed and its RCoA remains the same. It then follows that a mobile node actually only needs a new reservation path between its new AR and the MAP, and maintains the same reservation path between the MAP and its correspondent node. In the proposed schemes, the MAP is configured so that it can assist the mobile node to make this kind of partial resource reservation: With the information on the current binding between the mobile node's LCoA and RCoA, the MAP is capable of intercepting and looking into the RSVP messages and swapping the LCoA and RCoA of the packets in a way that the reservation paths below and above the MAP are identified by the LCoA and the RCoA respectively.[I] As a result, as long as the mobile node moves within the same MAP domain, the RSVP signalling only traverses the network up to the MAP and the reservation re-establishment signalling procedure is optimised.

There are also schemes such as [84–86] for MIPv4 that do not re-establish reservation paths between the mobile node and its correspondent node as normal. These schemes build a *forwarding chain* using an RSVP tunnel [89] between the mobile node's current and previous ARs, and require all the data and signalling packets to be transmitted through the tunnel. Therefore, the previously established RSVP

[I]The basic RSVP in this case cannot achieve end-to-end resource reservation within the MAP domain since RSVP signalling messages are encapsulated and thus are transparent to RSVP-compliant routers.

reservation path can still be used and only a new segment of RSVP reservation path is added. However, this type of scheme has the problem of routing inefficiency since the traffic is forced to travel through all the mobile node's previous ARs.

4.3 PEPA Models of the Basic and Mobility-supported RSVP

In this section, the PEPA models of the basic and mobility-supported RSVP are presented. Here the mobility-supported RSVP refers to the schemes that optimise the procedure of re-establishing an RSVP reservation path between the two communicating ends after a handover.

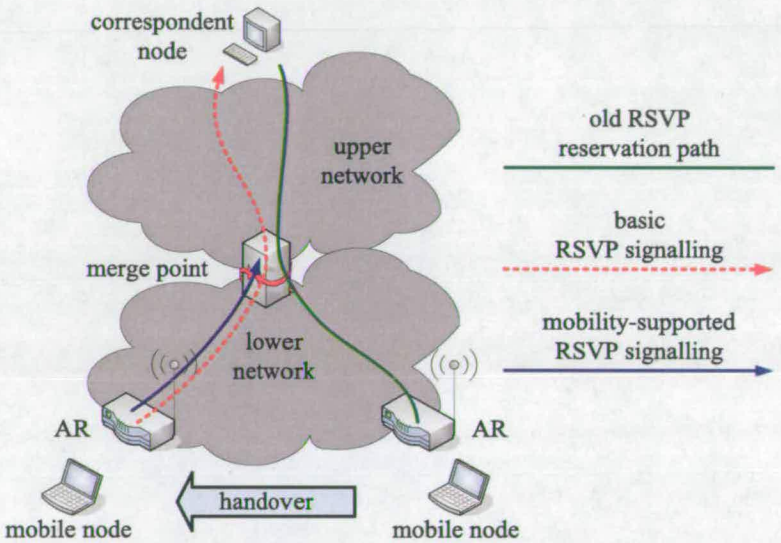


Figure 4.1: *Different reservation request procedures*

Figure 4.1 shows the reservation request procedures of the basic and mobility-supported RSVP. A mobile node engages in a QoS session with its correspondent node and preserves an (old) RSVP reservation path between them. During the communication, it may perform handovers and need to re-establish a new RSVP reservation path. In the models the old and new RSVP reservation paths are assumed to have a shared part, e.g. in HMIPv6 a mobile node performs a handover within its MAP domain. In this situation, the signalling of the mobility-supported RSVP only traverses the segment of the network to the merge point of the paths whereas that of the basic RSVP traverses the whole network up to the correspondent node.

The network between the communicating ends can be separated into two parts according to the merge point of the old and new RSVP reservation paths, i.e., the

network below and above the merge point are referred to as the *lower network* and the *upper network* respectively. The lower network usually consists of the whole or part of the mobile node's access network (depending on the position of the merge point) and the upper network generally consists of the Internet core network which is typically heavily loaded. Moreover, the concept of a *channel* is borrowed from WWANs, and resources of the lower and upper networks are represented as *lower network channels* and *upper network channels* respectively. In the following PEPA models, three types of PEPA components are defined to model the mobile node and the resources of the lower and upper networks, namely MN, $CHAN^L$ and $CHAN^U$ respectively. There is another PEPA component CT defined to model the *cleanup timer* that is required by RSVP.

4.3.1 Traffic and Mobility Models

Before the PEPA models are built, it is necessary to make assumptions about the traffic and mobility models of a mobile node. The traffic model of a mobile node usually consists of two parts: the session arrival rate and the session duration. Although recent study suggests that the Internet traffic at the packet level exhibits long-range dependence (LRD), the Poisson process is still a good model of the session arrival behaviour of the mobile node [90, 91]. As for the session duration, the traditional exponential distribution is used [92–94].

The mobility model is generally concerned with the distribution of the residence time of a mobile node in an area. By using different assumptions about the speed, direction and area shape, various types of distributions can be derived. However, the handover behaviour of a mobile node also depends on the type of handover procedure and thus the distributions based on contrived mobility patterns and area shapes are not really practical [95]. For that reason, without any proved probability distribution, the exponential distribution is chosen to model the mobile node's residence time in an area.

4.3.2 PEPA Model of the Basic RSVP

Mobile Node: The mobile node is initially in the idle state MN_{Idle}. It generates new session requests (activity *session_arrive*) at the rate of λ and then goes to state $MN^{New}_{Request}$, in which it tries to make reservation paths in both lower and upper networks (activity $reserve_{new}$). If there are resources available in both lower and upper

networks, the request is accepted and the mobile node goes to state $MN_{Engaged}$ and carries out its session (activity *session*). Otherwise, the request could be blocked by either the lower network or the upper network (activities $block_{lower}$ and $block_{upper}$) and the mobile node keeps requesting until it is finally admitted by both networks. The mean duration of an session is assumed to be $1/\mu$. During the session, the mobile node may perform a handover (activity *handover*) at the frequency of υ, and it goes to state $MN^{Handover}_{Request}$ and requests new reservation paths in both lower and upper networks (activity $reserve^{all}_{handover}$) in order to continue its session.[II] After the session is completed, the mobile node goes to state $MN_{Completed}$, tears down its current reservation path (activity *tear_down*), and goes back to the idle state. The component MN is defined as:

$$
\begin{aligned}
MN_{Idle} &\stackrel{def}{=} (session\_arrive,\ \lambda).MN^{New}_{Request} \\
MN^{New}_{Request} &\stackrel{def}{=} (reserve_{new},\ r_n).MN_{Engaged} \\
&\quad + (block_{lower},\ r^l_b).MN^{New}_{Request} \\
&\quad + (block_{upper},\ r^u_b).MN^{New}_{Request} \\
MN_{Engaged} &\stackrel{def}{=} (session,\ \mu).MN_{Completed} \\
&\quad + (handover,\ \upsilon).MN^{Handover}_{Request} \\
MN^{Handover}_{Request} &\stackrel{def}{=} (reserve^{all}_{handover},\ r^a_h).MN_{Engaged} \\
&\quad + (block_{lower},\ r^l_b).MN^{Handover}_{Request} \\
&\quad + (block_{upper},\ r^u_b).MN^{Handover}_{Request} \\
MN_{Completed} &\stackrel{def}{=} (tear\_down,\ r_t).MN_{Idle}
\end{aligned}
$$

As an example, the state transition diagram of the component MN is shown in Figure 4.2. The state transition diagrams of all the PEPA components in this thesis can be produced in a similar way.

Lower Network Channel: The component $CHAN^L$ models the resources in the lower network. It can be reserved (activities $reserve_{new}$ and $reserve^{all}_{handover}$) and torn down (activity *tear_down*) explicitly by a mobile node. Therefore, the definition of $CHAN^L$ is similar to the behaviour of a simple queue in queueing theory. The state of the lower network is denoted as $CHAN^L_i$, where the subscript i indicates the number of engaged channels. When the mobile node performs a handover, its

[II]The mobile node is assumed to implement the *local repair* [24] option, so it can send reservation requests almost immediately after a handover.

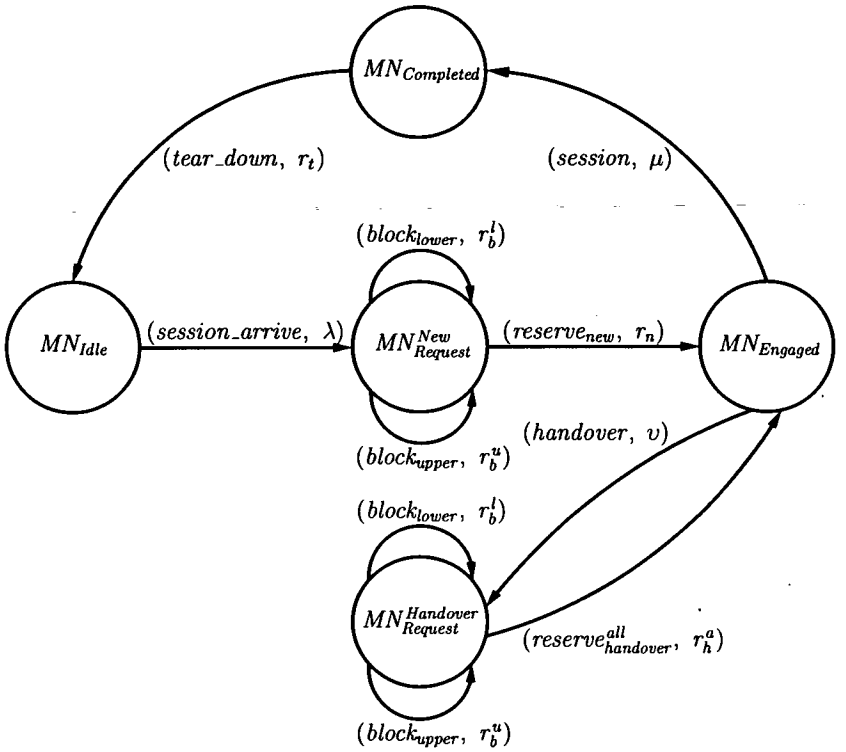


Figure 4.2: *State transition diagram of the PEPA component for mobile nodes*

old reservation path is still valid until the cleanup timer expires (activity *expire*).[III] When the $CHAN^L$ is fully engaged, it blocks the requests of mobile nodes (activity $block_{lower}$). If the capacity of the lower network is M, the component $CHAN^L$ is defined as ($i = 1, 2, \cdots, M-1$):

$$
\begin{aligned}
CHAN_0^L &\stackrel{def}{=} (reserve_{new},\ \top).CHAN_1^L \\
&\quad + (reserve^{all}_{handover},\ \top).CHAN_1^L \\
CHAN_i^L &\stackrel{def}{=} (reserve_{new},\ \top).CHAN_{i+1}^L \\
&\quad + (reserve^{all}_{handover},\ \top).CHAN_{i+1}^L \\
&\quad + (tear\_down,\ \top).CHAN_{i-1}^L \\
&\quad + (expire,\ \top).CHAN_{i-1}^L \\
CHAN_M^L &\stackrel{def}{=} (block_{lower},\ \top).CHAN_M^L \\
&\quad + (tear\_down,\ \top).CHAN_{M-1}^L \\
&\quad + (expire,\ \top).CHAN_{M-1}^L
\end{aligned}
$$

[III]The basic RSVP [24] only requires a node to explicitly tear down its reservation path at the end of its session.

Upper Network Channel: The component $CHAN^U$ models the resources in the upper network and its behaviour is the same as that of $CHAN^L$. If the capacity of the upper network is N, the component $CHAN^U$ is defined as ($i = 1, 2, \cdots, N-1$):

$$\begin{aligned}
CHAN_0^U &\stackrel{def}{=} (reserve_{new},\ \top).CHAN_1^U \\
&\quad + (reserve_{handover}^{all},\ \top).CHAN_1^U \\
CHAN_i^U &\stackrel{def}{=} (reserve_{new},\ \top).CHAN_{i+1}^U \\
&\quad + (reserve_{handover}^{all},\ \top).CHAN_{i+1}^U \\
&\quad + (tear_down,\ \top).CHAN_{i-1}^U \\
&\quad + (expire,\ \top).CHAN_{i-1}^U \\
CHAN_N^U &\stackrel{def}{=} (block_{upper},\ \top).CHAN_N^U \\
&\quad + (tear_down,\ \top).CHAN_{N-1}^U \\
&\quad + (expire,\ \top).CHAN_{N-1}^U
\end{aligned}$$

Cleanup Timer: The component CT models the cleanup timer that is used by RSVP to maintain the reserved resources and implement *soft reservation*. It is defined in such a way that it starts a timer for each obsolete reservation path caused by the mobile node's handover, and the reservation path is not released unless the timer expires. Each timer has a mean lifetime of $1/\tau$ (activity *expire*). Therefore, after i *consecutive* handovers, there are i timers enabled and the aggregate expiration rate of all the obsolete reservation paths is $i * \tau$. Moreover, since the mobile node always requires reservation paths in the lower and upper networks together, the maximum number of reservation paths in the whole network is $\min(M, N)$, which is also the maximum number of enabled timers. The component CT is defined as ($i = 1, 2, \cdots, \min(M, N) - 1$):

$$\begin{aligned}
CT_0 &\stackrel{def}{=} (handover,\ \top).CT_1 \\
CT_i &\stackrel{def}{=} (handover,\ \top).CT_{i+1} \\
&\quad + (expire,\ i * \tau).CT_{i-1} \\
CT_{\min(M,N)} &\stackrel{def}{=} (expire,\ \min(M, N) * \top).CT_{\min(M,N)-1}
\end{aligned}$$

System Definition: Since the mobile node using the basic RSVP reserves the resources of the whole network for both new and handover sessions, the activities $reserve_{new}$ and $reserve_{handover}^{all}$ must be synchronised by MN, $CHAN^L$ and $CHAN^U$. $CHAN^L$ and $CHAN^U$ can cooperate with MN on the activities $block_{lower}$ and $block_{upper}$

respectively when they are fully engaged. The CT synchronises with MN on the *handover* activity and with $CHAN^L$ and $CHAN^U$ on the *expire* activity. When the mobile node finishes the session, the activity *tear_down* is carried out by MN, $CHAN^L$ and $CHAN^U$ together. The PEPA model of the basic RSVP is constructed as (the number K in the square brackets denotes K parallel mobile nodes):

$$RSVP^B \stackrel{def}{=} \Big(MN_{Idle}[K]\Big) \underset{L_1}{\bowtie} \Big(\Big(CHAN_0^L \underset{L_2}{\bowtie} CHAN_0^U\Big) \underset{L_3}{\bowtie} CT_0\Big),$$

where

$$L_1 = \Big\{reserve_{new}, reserve^{all}_{handover}, tear\_down, block_{lower}, block_{upper}, handover\Big\},$$
$$L_2 = \Big\{reserve_{new}, reserve^{all}_{handover}, tear\_down, expire\Big\}, \text{ and } L_3 = \Big\{expire\Big\}.$$

4.3.3 PEPA Model of the Mobility-supported RSVP

Mobile Node: Unlike the mobile node using the basic RSVP, the mobile node using the mobility-supported RSVP only requires a new reservation path in the lower network after a handover. This is implemented by changing the behaviour of the component MN after a handover. That is, in state $MN^{Handover}_{Request}$ the component MN performs the activity $reserve^{lower}_{handover}$. In this case, the mobile node can only be blocked by the lower network (activity $block_{lower}$). Therefore, the component MN is defined as:

$$\begin{aligned}
MN_{Idle} &\stackrel{def}{=} (session_arrive,\ \lambda).MN^{New}_{Request}\\
MN^{New}_{Request} &\stackrel{def}{=} (reserve_{new},\ r_n).MN_{Engaged}\\
&\quad + (block_{lower},\ r_b^l).MN^{New}_{Request}\\
&\quad + (block_{upper},\ r_b^u).MN^{New}_{Request}\\
MN_{Engaged} &\stackrel{def}{=} (session,\ \mu).MN_{Completed}\\
&\quad + (handover,\ \upsilon).MN^{Handover}_{Request}\\
MN^{Handover}_{Request} &\stackrel{def}{=} (reserve^{lower}_{handover},\ r_h^l).MN_{Engaged}\\
&\quad + (block_{lower},\ r_b^l).MN^{Handover}_{Request}\\
MN_{Completed} &\stackrel{def}{=} (tear_down,\ r_t).MN_{Idle}
\end{aligned}$$

Lower Network Channel: The component $CHAN^L$ is modified so that it is aware of the new type of request which only reserves the resources of the lower network. Generally, the mobility-supported RSVP requires the mobile node to explicitly tear

down its obsolete reservation path after a handover. However, to avoid a one-sided comparison, this requirement is not implemented. The component $CHAN^L$ is modified as ($i = 1, 2, \cdots, M-1$):

$$\begin{aligned}
CHAN_0^L &\stackrel{def}{=} (reserve_{new}, \top).CHAN_1^L \\
&\quad + (reserve^{lower}_{handover}, \top).CHAN_1^L \\
CHAN_i^L &\stackrel{def}{=} (reserve_{new}, \top).CHAN_{i+1}^L \\
&\quad + (reserve^{lower}_{handover}, \top).CHAN_{i+1}^L \\
&\quad + (tear_down, \top).CHAN_{i-1}^L \\
&\quad + (expire, \top).CHAN_{i-1}^L \\
CHAN_M^L &\stackrel{def}{=} (block_{lower}, \top).CHAN_M^L \\
&\quad + (tear_down, \top).CHAN_{M-1}^L \\
&\quad + (expire, \top).CHAN_{M-1}^L
\end{aligned}$$

Upper Network Channel: In the mobility-supported RSVP, there is no need for a new reservation path in the upper network after a handover. An upper network reservation path is only established and torn down at the start and the end of an session, and it is not aware of the mobile node's handover and consequently never expires. The component $CHAN^U$ is defined as ($i = 1, 2, \cdots, N-1$):

$$\begin{aligned}
CHAN_0^U &\stackrel{def}{=} (reserve_{new}, \top).CHAN_1^U \\
CHAN_i^U &\stackrel{def}{=} (reserve_{new}, \top).CHAN_{i+1}^U \\
&\quad + (tear_down, \top).CHAN_{i-1}^U \\
CHAN_N^U &\stackrel{def}{=} (block_{upper}, \top).CHAN_N^U \\
&\quad + (tear_down, \top).CHAN_{N-1}^U
\end{aligned}$$

Cleanup Timer: The component CT in the mobility-supported RSVP has the same function as that in the basic RSVP. The only difference is that the maximum number of timers in the whole network is M because only lower network reservation paths expire and the mobile node can use up all the resources in the lower network. The

component CT is defined as ($i = 1, 2, \cdots, M-1$):

$$\begin{aligned} CT_0 &\stackrel{def}{=} (handover,\ \top).CT_1 \\ CT_i &\stackrel{def}{=} (handover,\ \top).CT_{i+1} \\ &\quad + (expire,\ i * \tau).CT_{i-1} \\ CT_M &\stackrel{def}{=} (expire,\ M * \top).CT_{M-1} \end{aligned}$$

System Definition: The system definition of the mobility-supported RSVP is the same as that of the basic RSVP, except for the cooperation sets L_1 and L_2. In L_1 the activity $reserve^{all}_{handover}$ is replaced by the activity $reserve^{lower}_{handover}$. In L_2 there are no longer activities $reserve^{all}_{handover}$ and $expire$, since there is no need to establish an upper network reservation path after a handover and the reservation paths in the upper network do not expire. The PEPA model of the mobility-supported RSVP is constructed as:

$$RSVP^{MS} \stackrel{def}{=} \Big(MN_{Idle}[K]\Big) \underset{L_1}{\bowtie} \Big(\Big(CHAN^L_0 \underset{L_2}{\bowtie} CHAN^U_0\Big) \underset{L_3}{\bowtie} CT_0\Big),$$

where

$$L_1 = \Big\{reserve_{new}, reserve^{lower}_{handover}, tear\_down, block_{lower}, block_{upper}, handover\Big\},$$
$$L_2 = \Big\{reserve_{new}, tear\_down\Big\}, \text{ and } L_3 = \Big\{expire\Big\}.$$

4.3.4 System States of the PEPA Models

Due to the characteristics of the basic and mobility-supported RSVP, resources of the lower and upper networks are consumed in different ways. For the basic RSVP, since a mobile node requires resources all through the network for both new and handover sessions, there are at most $\min(M, N)$ reservation paths in both lower and upper networks. For the mobility-supported RSVP, since upper network resources are only reserved for new sessions, there are at most $\min(K, N)$ reservation paths in the upper network.[IV] On the other hand, all the M channels of the lower network can be used for new and handover sessions. For example, if $K = 15$, $M = 20$ and $N = 10$, then for the basic RSVP, the upper network could be fully engaged whereas 10 channels in the lower network are redundant. For the mobility-supported RSVP,

[IV] Remember that in both PEPA models, a mobile node cannot start a new session until it finishes its current session.

both networks could be fully engaged.

To ensure that the models are numerically tractable, in both models there are $K = 5$ parallel mobile nodes and M and N are set to 7 and 5 respectively. The capacity of the upper network in particular is set smaller than that of the lower network to emphasise the scarcity of resources in the upper network. The system states of the basic RSVP model and the mobility-supported RSVP model are the *feasible* combinations of the state of each component of that model. External codes written in AWK [96], which is designed for processing text-based data, are used to check whether every state is feasible. For example, for the basic RSVP model, the number of engaged upper network channels, the number of engaged lower network channels and that of engaged mobile nodes should be the same; and for the mobility-supported RSVP model, the number of engaged upper network channels should equal to that of engaged mobile nodes.

In this thesis, the system state of a model is denoted as a *tuple*. For instance, both models have the following system state:

$$\left(MN_{Request}^{Handover} | MN_{Engaged} | MN_{Idle} | MN_{Idle} | CHAN_2^L | CHAN_2^U | CT_1\right),$$

where the symbol "|" has no operational meaning but just separates different components.

4.4 Performance Evaluation

Since the network between the two communicating ends usually consists of the Internet core network where the traffic is typically congested, an optimised utilisation of it is both practically and economically required. A more congested network usually results in a higher reservation request blocking probability, and a larger signalling delay implies a longer interruption of QoS sensitive traffic. The mobility-supported RSVP is especially designed to eliminate unnecessary consumption of network resources and reduce the signalling overheads. Therefore, the performance measures investigated in this work are the probability of rejecting handover reservation requests and the costs associated with the handover reservation signalling.

Rate	Corresponding activity	Value (second^{-1})
λ	$session\_arrive$	$[10 : 10 : 150].^{-1}$
μ	$session$	$[30 : 30 : 450].^{-1}$
υ	$handover$	$[60 : 60 : 900].^{-1}$
τ	$expire$	150^{-1}
r_n	$reserve_{new}$	0.1^{-1}
r_h^a	$reserve_{handover}^{all}$	0.1^{-1}
r_h^l	$reserve_{handover}^{lower}$	0.05^{-1}
r_b^l	$block_{lower}$	0.025^{-1}
r_b^u	$block_{upper}$	0.075^{-1}
r_t	$tear\_down$	0.1^{-1}

Table 4.1: *Activity rates of the PEPA models of the basic and mobility-supported RSVP*

4.4.1 Parameter Settings

The activity rates used in the basic RSVP model and the mobility-supported RSVP model are set as follows. The session arrival interval $1/\lambda$, the session duration $1/\mu$, and the mobile node's handover interval $1/\upsilon$ are the control parameters and their variation ranges are listed in Table 4.1.[v] The mean lifetime of an RSVP reservation path $1/\tau$ is set to 150 seconds as suggested in [24]. The mean delays of the signalling messages that traverse the whole and the lower network are set to 0.1s and 0.05s respectively. As for the lower and upper network reservation request blocking signalling messages, since they could be generated at any RSVP-compliant router in the lower and upper networks, it is assumed that statistically they are generated at the middle point of the network and their mean delays are set to 0.025s and 0.075s respectively. The rates of all the activities are listed in Table 4.1.

4.4.2 Handover Blocking Probability

Since the mobile nodes in the basic RSVP model and the mobility-supported RSVP model are individually expressed, performance measures can be derived by observing either a single mobile node or all the mobile nodes as a whole (the former is chosen). Moreover, as the 4 *MN* components in both models have identical behaviour, any one of them can be chosen to carry out the evaluation (the first one is chosen).

[v] The operator ":" generates a sequential series of numbers, each number separated by a step value, using the syntax first:step:last; the operator ".$^{-1}$" calculates the reverse of each element of a matrix.

The handover blocking happens in the system states in which the mobile node requires a reservation path after a handover whilst the lower network or the upper network is fully engaged. Therefore, in the basic RSVP model these system states can be described as the union of two sets of system states:

$$S_{HB}^{B} = \left\{ \left(MN_{Request}^{Handover} | * | * | * | CHAN_{M}^{L} | * | * \right) \right\} \bigcup \left\{ \left(MN_{Request}^{Handover} | * | * | * | * | CHAN_{N}^{U} | * \right) \right\},$$

and in the mobility-supported RSVP model these system states can be described as the set:

$$S_{HB}^{M} = \left\{ \left(MN_{Request}^{Handover} | * | * | * | CHAN_{M}^{L} | * | * \right) \right\},$$

where the symbol "$*$" is used as a wildcard character for a component's state and means that there is no requirement on the state of that component as long as S_{HB}^{B} and S_{HB}^{M} are the subsets of the state space of their respective models. In these system states, the mobile node can only perform the self-transitions corresponding to blocking until there are available resources. Therefore, the percentage of time that the mobile node gets stuck due to the lack of resources, which can also be regarded as the handover blocking probability experienced by the mobile node, is calculated as:

$$P_{HB}^{B} = \sum_{s_i \in S_{HB}^{B}} \pi(s_i), \tag{4.1}$$

$$P_{HB}^{M} = \sum_{s_i \in S_{HB}^{M}} \pi(s_i), \tag{4.2}$$

where $\pi(s_i)$ is the equilibrium probability of the system state s_i, and P_{HB}^{B} and P_{HB}^{M} denote the handover blocking probability for the basic RSVP model and the mobility-supported RSVP model respectively.

In the following subsections, the effect of session arrival interval, session duration and handover interval on the handover blocking probability is investigated.

4.4.2.1 Effect of Session Arrival Interval

Figure 4.3 shows the effect of the session arrival rate on the handover blocking probability.[VI] The handover interval is set to 540 seconds and the session duration

[VI] This figure is produced by the following steps: For each value of the session arrival interval, firstly, the equilibrium probability vector of each PEPA model is derived using Eq. (3.1) and Eq. (3.2), and then their handover blocking probabilities are calculated using Eq. (4.1) and Eq. (4.2). In this thesis, all the figures showing evaluation results are produced in a similar way: For each value of a control parameter,

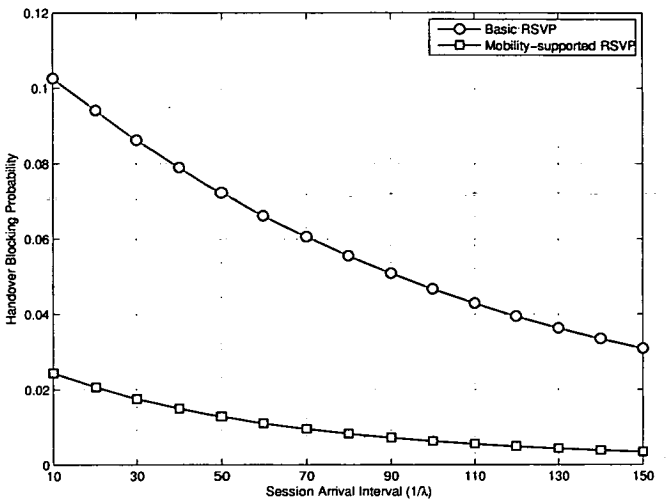


Figure 4.3: *The effect of session arrival interval on handover blocking probability*

is set to 240 seconds. It can be observed that as the interval between session arrivals grows, the blocking probability for the basic RSVP decreases from 1.03×10^{-1} to 3.09×10^{-2}, and the mobility-supported RSVP has a lower blocking probability ranging from 2.44×10^{-2} to 3.48×10^{-3} because it does not require a new reservation path in the upper network after a handover. Moreover, the mobility-supported RSVP has a much smaller absolute change than the basic RSVP on the handover blocking probability and their difference is larger at smaller session arrival intervals. This indicates that the mobility-supported RSVP is less affected by the traffic intensity and performs better when sessions are generated frequently.

4.4.2.2 Effect of Session Duration

Figure 4.4 shows how the session duration affects the handover blocking probability. The results are calculated by setting the handover interval and the session arrival interval to be 540 and 80 seconds respectively. It can be observed that the handover blocking probability for both schemes increases as the session duration becomes longer. When the duration is 30 seconds, the basic RSVP has a blocking probability of 4.72×10^{-4} whereas the mobility-supported RSVP has a blocking probability of 1.24×10^{-5}, which means that the mobile node is highly unlikely to be blocked for handover if it uses mobility-supported RSVP. When the session duration grows to 450

the equilibrium probability vector of a PEPA model is first derived, and the performance measure of interest is then calculated using the corresponding equation.

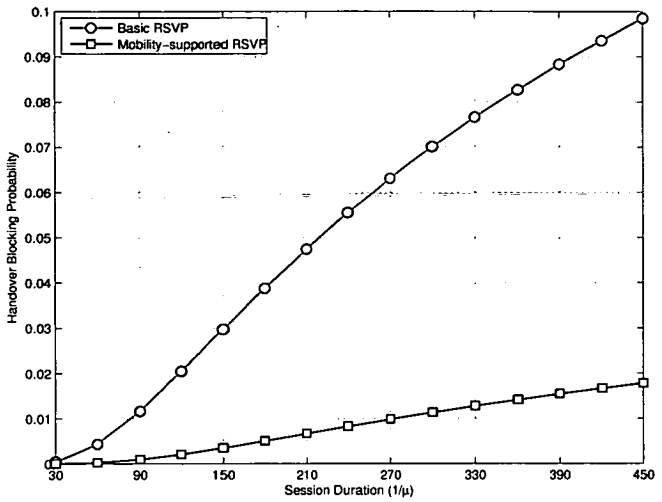


Figure 4.4: *The effect of session duration on handover blocking probability*

seconds, the mobility-supported RSVP has an increase in blocking probability up to 1.78×10^{-2}, whereas that of the basic RSVP reaches 9.85×10^{-2}. This again indicates that at higher traffic intensity the mobility-supported RSVP performs better than the basic RSVP.

4.4.2.3 Effect of Handover Interval

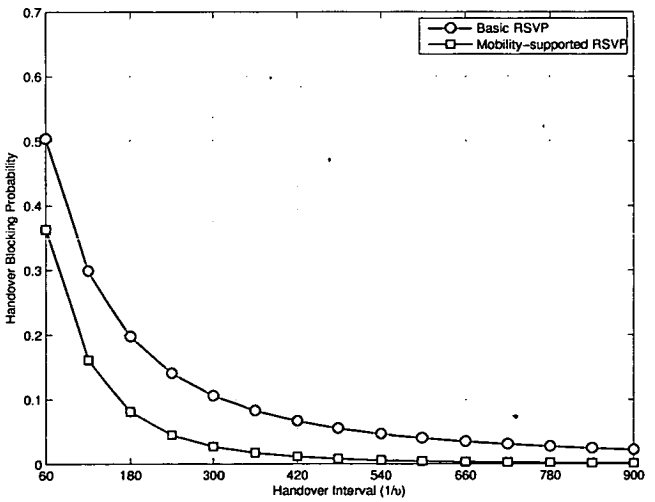


Figure 4.5: *The effect of handover interval on handover blocking probability*

The effect of the mobile node's mobility is also investigated, and the results are shown in Figure 4.5. The session arrival interval and session duration are set to 80

and 240 seconds respectively. It is clear that when the handover frequency decreases, the reservation requests for network resources are reduced, thereby resulting in a lower blocking probability. For the basic RSVP, there is a recognisable decrease in the blocking probability from 5.04×10^{-1} to 2.27×10^{-2}. On the other hand, the blocking probability for the mobility-supported RSVP ranges from 3.63×10^{-1} to 1.48×10^{-3}, and remains almost unchanged for handover intervals larger than 720 seconds. Moreover, the difference in performance of the two schemes diverges at small handover intervals, which means the mobility-supported RSVP can significantly improve the performance in high mobility scenarios.

4.4.3 Handover Signalling Cost

Since one of the major benefits of the mobility-supported RSVP is reducing the scope of the network over which the RSVP signalling traverses after a handover, another performance measure of interest is the handover signalling cost. The set of system states in which the mobile node can perform handover is denoted as S_{HS}. In the basic RSVP model, the handover signalling activity is $reserve^{all}_{handover}$ and S_{HS} is the union of the sets described as:

$$S^B_{HS} = \bigcup_{\substack{0 \le i \le M-1 \\ 0 \le j \le N-1}} \left\{ \left\{ \left(MN^{Handover}_{Request} | * | * | * | CHAN^L_i | CHAN^U_j | * \right) \right\} \right\},$$

and in the mobility-supported RSVP model, the handover signalling activity is $reserve^{lower}_{handover}$ and S_{HS} is the union of the sets described as:

$$S^M_{HS} = \bigcup_{0 \le i \le M-1} \left\{ \left\{ \left(MN^{Handover}_{Request} | * | * | * | CHAN^L_i | * | * \right) \right\} \right\},$$

where S^B_{HS} and S^M_{HS} should be the subsets of the state space of their respective models.

By employing MRS and Eq. (3.3), the handover signalling costs for the basic RSVP model (C^B_{HS}) and the mobility-supported RSVP model (C^M_{HS}) are calculated as:

$$C^B_{HS} = \sum_{s_i \in S^B_{HS}} r^B_i * \pi(s_i), \tag{4.3}$$

$$C^M_{HS} = \sum_{s_i \in S^M_{HS}} r^M_i * \pi(s_i), \tag{4.4}$$

where r_i^B and r_i^M denote the rewards associated with the activities $reserve_{handover}^{all}$ and $reserve_{handover}^{lower}$ in the system state s_i respectively. Moreover, the rewards for the same type of handover signalling activity in each system state are set to equal values. Therefore, Eq. (4.3) and Eq. (4.4) can be written as:

$$C_{HS}^B = r^B * \sum_{s_i \in S_{HS}^B} \pi(s_i) = r^B * P_{HS}^B, \tag{4.5}$$

$$C_{HS}^M = r^M * \sum_{s_i \in S_{HS}^M} \pi(s_i) = r^M * P_{HS}^M. \tag{4.6}$$

It is necessary to point out here that the summations $\sum_{s_i \in S_{HS}^B} \pi(s_i)$ and $\sum_{s_i \in S_{HS}^M} \pi(s_i)$ (denoted as P_{HS}^B and P_{HS}^M respectively) are in fact the percentage of time the mobile node spends on the handover signalling activities in the respective models, and their values are in inverse proportion to the rates of the handover signalling activities.

The definition of the rewards r^B and r^M is *not* unique. Since the effect of the signalling *delay* is already reflected in the summations in Eq. (4.5) and Eq. (4.6), the same reward r is used for both types of handover signalling activities and it can be regarded as the *cost* (e.g. packet loss rate) associated with the handover signalling. The value of r will be discussed in Section 4.4.3.1. As in Section 4.4.2, the effect of the traffic intensity and the mobility pattern on the handover signalling cost is investigated, and the activity rates for each of the following subsections are the same as their counterpart in Section 4.4.2.

4.4.3.1 Effect of Session Arrival Interval

The effect of the session arrival interval on the handover signalling cost is shown in Figure 4.6. The reward r is set to the inverse of the minimum of the P_{HS}^B and P_{HS}^M values calculated from the session arrival intervals. Equivalently, the reward r makes the minimum of the calculated C_{HS}^B and C_{HS}^M equal 1. The reward is designed in this way so as to normalise the handover signalling cost. In Figure 4.6, decreases (nearly linear) in the handover signalling costs for both schemes can be observed. This is because at smaller session arrival intervals, the mobile node engages in communications more frequently and therefore is more likely to perform handovers, which increases the cost of handover signalling. Within the variation range of the session arrival interval, the handover signalling cost for the basic RSVP (from 2.46 to 1.90) is roughly twice as much as that for the mobility-supported RSVP (from 1.50 to 1.00), because the

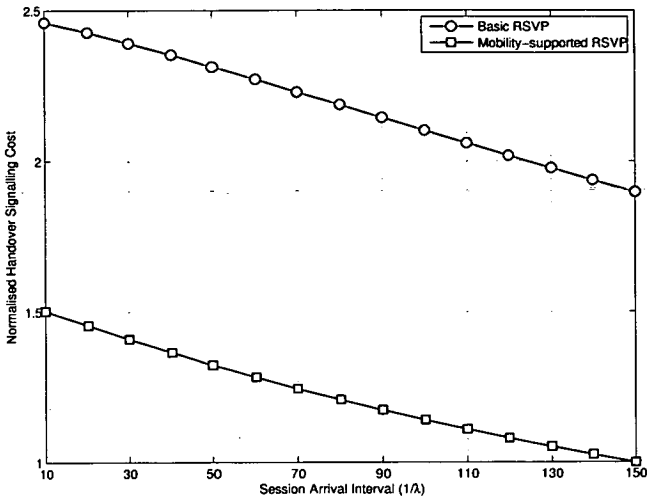


Figure 4.6: *The effect of session arrival interval on handover signalling cost*

latter restricts the RSVP signalling to be within the affected area of the network. Moreover, their difference gets larger for small intervals, which suggests that the mobility-supported RSVP reduces handover signalling cost at high traffic intensity scenarios.

4.4.3.2 Effect of Session Duration

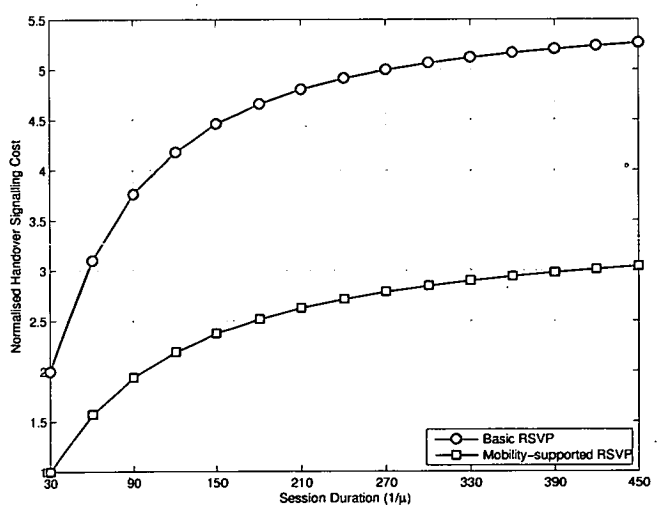


Figure 4.7: *The effect of session duration on handover signalling cost*

Figure 4.7 shows how the session duration affects the handover signalling cost. As the session duration increases, the handover signalling costs for both schemes grow

approximately threefold, which is because a longer session implies more handovers for the mobile node. The cost for the basic RSVP grows from around 2.00 to 5.27, and the cost for the mobility-supported RSVP is approximately half of that for the basic RSVP (from 1.00 to 3.04). Their difference gets larger for longer session durations, which again indicates the benefits of the mobility-supported RSVP.

4.4.3.3 Effect of Handover Interval

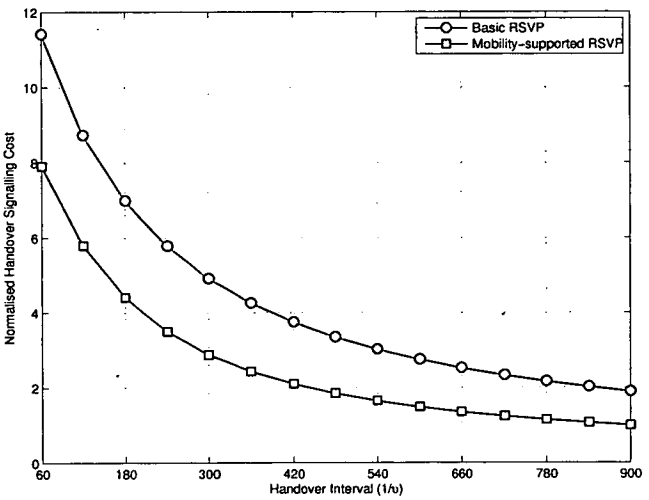


Figure 4.8: *The effect of handover interval on handover signalling cost*

Results similar to those in the previous two subsections can also be identified in Figure 4.8, where the control parameter is the handover interval. The mobility-supported RSVP experiences a lower signalling cost than the basic RSVP, and clearly, the mobile node can save its handover signalling cost by avoiding changing its connectivity too often. For the basic RSVP, the handover signalling cost ranges from around from 11.42 to 1.90, and that for the mobility-supported RSVP ranges from 7.91 to 1.00. For the large handover intervals, the difference between the two schemes is smaller because the mobile node seldom changes its point of attachment and the benefits of the mobility-supported RSVP is less apparent. However, the mobility-supported RSVP has a greater improvement at small handover intervals, which again shows that it performs better in high mobility scenarios.

4.5 Conclusions

Since RSVP and mobility management protocols were designed independently, their efficient integration is necessary in order to provide a QoS guaranteed mobility in mobile wireless networks. One problem of this integration is the signalling optimisation problem which results from the reservation re-establishment after a mobile node performs a handover. In this chapter, PEPA models of the basic and mobility-supported RSVP have been constructed and evaluated in order to verify the advantages of the signalling optimisation schemes. The effect of the traffic and mobility patterns of mobile nodes on the performance of the basic and mobility-supported RSVP have been compared. Two performance measures, the probability of blocking RSVP reservation requests after a handover and the cost of RSVP signalling, have been investigated because they determine the QoS perceived by the users. The results indicate that the mobility-supported RSVP clearly outperforms the basic RSVP on both measures, and verify that the former is more suitable in high traffic and mobility scenarios. These enhancements are achieved by avoiding unnecessary resource reservation paths in the unaffected part of the network and limiting RSVP signalling within the affected part. Moreover, both measures are much more sensitive to the mobility pattern than to the traffic pattern, i.e., the mobility of the mobile node has the largest effect on its handover performance. However, despite the fact that the mobility-supported RSVP can reduce the interruptions to a mobile node during its handover, an interruption can still happen, especially when there are not enough resources at the mobile node's new location for it to use. Therefore, schemes that reserve resources in advance have been proposed and they can further reduce handover interruptions that happen to the mobile node. The modelling of these schemes will be studied in the next chapter.

It is important to point out that although only models of small size are used for performance evaluation, the presented PEPA models have no restriction on model size. Certainly, to obtain more realistic results, larger scale models consist of more components should be investigated. Large-size PEPA models can be solved using a set of coupled ordinary differential equations (ODEs) and meaningful performance measures such as utilisation and response time can be derived directly without generating the whole state spaces of the models [97]. However, since the performance measures investgated in this work are related to the joint probability of multiple

components, the derivation of these measures from the ODEs is not easy and is still under investigation. Although the PEPA models in this chapter are not complex and the results derived from them are rather predictable, the characteristics of the modelled schemes have been captured by the models and they are expressed in terms of the behaviour of network entities. In the following chapters, more sophisticated PEPA models will be presented.

Chapter 5
Modelling of Advance Resource Reservation Schemes for RSVP in Mobile Wireless Networks

Another major problem of deploying RSVP in mobile wireless networks is called the advance resource reservation (ARR) problem with regard to reserving resources in advance for a mobile node. Traditional solutions to this problem waste too many network resources and increase the probability of blocking active reservation requests. In this chapter, a novel ARR scheme which optimises network resource utilisation is proposed and the PEPA models of the conventional and proposed schemes are built and evaluated.

5.1 Introduction

As discussed in Chapter 4, resource reservation using conventional RSVP exhibits deficiencies in mobile wireless networks. This is because a mobile node must re-establish a reservation path after a network layer handover in order to continue its ongoing QoS session. When the mobile node changes the data flow path after a handover, the congestion levels at the routers along the old and new paths may also change [98]. If the new path is overcongested, the available bandwidth along the new path may not be sufficient to satisfy the requirements of the QoS session. Therefore, the mobile node may be rejected for making a reservation path and its QoS session will be interrupted. To solve this problem, it has been suggested that the required resources be reserved in advance in the networks that a mobile node may visit. In this way, resources are guaranteed before the handover, and the mobile node can continue its communication smoothly after it switches its network connectivity.

However, conventional ARR schemes waste network resources from the QoS traffic's perspective, and an efficient ARR scheme that can optimise network resource utilisation is needed. The contribution of the work presented in this chapter is

that a reservation optimised ARR scheme which combines resource reservation and call admission control (CAC) mechanisms is proposed. This scheme is designed to improve the performance of the existing ARR schemes by avoiding too many advance reservation paths in a network. Furthermore, this scheme takes account of the traffic and mobility patterns of mobile nodes and only allows the most eligible ones to reserve resources in advance. To demonstrate the advantages of the proposed scheme, PEPA models of different ARR schemes are built and assessed. These PEPA models are independent of the specific implementations of the ARR schemes, and important performance measures including the blocking probabilities of different types of reservation requests and the mean numbers of different types of reservation paths are derived.

The rest of this chapter is structured as follows. Section 5.2 gives a general review of some typical ARR schemes. In Section 5.3 a reservation optimised ARR scheme and its operating procedure are presented. In Section 5.4 the PEPA models of the different ARR schemes are described, and their performance is evaluated and compared in Section 5.5. In Section 5.6 the evaluation results are discussed.

5.2 Advance Resource Reservation Schemes for RSVP in Mobile Wireless Networks

ARR schemes aim to reduce the reservation request blocking probability during handover by reserving resources in advance for a mobile node. Previous proposals for ARR schemes can be classified into two types: *agent-based* and *multicast-based*.

- **Agent-based schemes**: In the agent-based schemes [99–102], there are two types of reservation paths: *active* and *passive*. A mobile node makes an active reservation path in its current network and passive reservation paths in neighbouring networks that it may visit. An active reservation path is actively used by the mobile node to carry out its communication and the passive reservation paths are just reserved for the mobile node but not used. When the mobile node hands over into a neighbouring network, its passive reservation path in the newly visited network is switched to the active state and its old active reservation path is changed to the passive state. In every network there is an agent which is in charge of the advance resource reservation and reservation state change procedures. The mobile node needs to inform

the agents of the neighbouring networks which it may visit of its reservation information and require them to make passive reservation paths for it. In [100], the passive reservation paths are only established when the mobile node is in the overlapped area of two networks and intends to perform a handover. To reduce redundant passive reservation paths, some approaches are equipped with the position prediction techniques [101, 102]. These techniques estimate the most likely neighbouring networks that the mobile node may visit according to statistical models or historical moving trajectories, and only allow advance resource reservations in the predicted networks.

- **Multicast-based schemes**: In the multicast-based schemes [82, 83, 103], RSVP signalling messages and ordinary data packets are delivered to a mobile node using IP multicast routing. As in the agent-based approaches, there is also an agent in every network. Before the mobile node starts its communication, the mobile node and the agents in its neighbouring networks join a multicast group and share a multicast address. All the traffic goes through the multicast address, and a handover of the mobile node can be regarded as leaving and joining the branches of a multicast tree. A *conventional* reservation path is established between the mobile node and its correspondent node, and the neighbouring agents make *predictive* reservation paths on behalf of the mobile node. These two types of reservation states are essentially the same as the active and passive states in the agent-based schemes. When the mobile node hands over into a neighbouring network, the states of its old and new reservation paths are changed accordingly. Since all the data packets are addressed to the multicast address, the mobile node can continue its communication without interruptions when it moves out of its current network.

With the help of the ARR schemes, the session interruptions during handover is reduced. However, an advance reservation path in a network is made exclusively for a certain mobile node and is not actively used by its reserver. Allowing too many advance reservation paths in a network will increase the blocking probability of the active reservation requests originating from the mobile nodes in that network. The approaches that allow traffic with lower QoS requirements to temporarily use the passive reservation paths [82,99,100,103] are not reliable, since the resources borrowed by a QoS session have to be returned when their reservers reclaim them, which results

in an interruption to the borrower. On the other hand, only allowing best-effort traffic to use the passive reservation paths is a waste of network resources from the QoS traffic's point of view. Therefore, putting a restriction on the number of the advance reservation paths in a network would be beneficial from the perspective of network utilisation. In fact, a combination of ARR and CAC mechanisms should achieve better performance on managing network resources as suggested in [98,104,105].

5.3 The Reservation Optimised Advance Resource Reservation Scheme

In this section, a reservation optimised ARR scheme is proposed. This scheme aims to achieve a better utilisation of the network resources by balancing the number of active and passive reservation paths in a network. The proposed scheme includes two admission mechanisms: passive reservation limited and session-to-mobility ratio (SMR) based replacement. It needs to be pointed out here that the proposed scheme is not a design of a new signalling procedure for an ARR scheme but is an investigation of an efficient way of utilising network resources.

5.3.1 Passive Reservation Limited Mechanism

In the conventional ARR schemes, the resources of a network are actively reserved by the mobile nodes in that network (namely local mobile nodes) and passively reserved by the mobile nodes in the neighbouring networks (namely foreign mobile nodes). Since the active and passive reservation requests are treated in the same way, there is no restriction on the number of passive reservation paths in a network. Moreover, since allowing too many passive reservation paths is a waste of network resources from the perspective of the QoS traffic, it is better to give higher priority to the active reservation requests because this type of request implies that there are QoS sessions that cannot start without the requested resources.

As in Chapter 4, the resources of a network are regarded as *channels*. To limit the number of passive reservation paths in a network, part of the channels are reserved only for active reservation paths. As a result, the channels of the network are partitioned into two groups: *standard channels* and *dedicated channels*. The only difference between a standard channel and a dedicated channel is that the former can be used for both active and passive reservation paths, whereas the latter can *only*

be used for active reservation paths. To guarantee that the channels are allocated correctly, there is an enhanced agent (EA) in each network which monitors the network resources and admits different types of reservation requests. The EA assigns dedicated channels or standard channels to the active reservation requests, whereas it only assigns standard channels to the passive reservation requests. In this way, the number of passive reservation paths in the network is limited and hence more resources are available for the active reservation requests. More importantly, unlike the conventional ARR schemes, the passive reservation limited mechanism does not allow an active reservation path to change to the passive state. That is, when the local mobile node hands over out of the local network, it has to release its active reservation path and instead requests a passive reservation path.

In order to avoid over-restricting passive reservation requests, the EA first allocates dedicated channels to the active reservation requests. The standard channels are only assigned when all the dedicated channels are engaged. Therefore, if the total number of channels in a network is T and the number of standard channels is S, then the maximum number of passive reservation paths in the network is S and the EA can accept at least $T - S$ active reservation requests.

5.3.2 SMR-based Replacement Mechanism

Since only the standard channels of a network can be used for passive reservation paths, they are scarce resources from the foreign mobile nodes' point of view. Consequently, an efficient admission strategy is necessary to determine which foreign mobile node is eligible to acquire a standard channel. Since the essential objective of an ARR scheme is to improve the handover performance of a mobile node, it would be better to assign a standard channel to the foreign mobile node which is most likely to handover during a session.[1]

In previous work such as [106, 107], the handover frequency of a mobile node is usually defined by the ratio of the mobile node's *session arrival rate* to its *mobility rate*. This type of ratio is used to optimise the packet routing and network traffic load in HMIPv6. However, a handover is the behaviour of a mobile node during its communication and it has no direct relationship with the session arrival rate. For that reason, in the proposed SMR-based replacement mechanism a modified form of the ratio, which is defined as the ratio of a mobile node's *session duration* to its *residence*

[1] All the sessions are assumed to have the same QoS class.

time in a network, is used to characterise the handover likelihood of the mobile node during its communication.

In the SMR-based replacement mechanism, every EA is assumed to be able to get the SMR information of the mobile nodes in its administrating network.[II] The mechanism works as follows: Assume there is a foreign mobile node which requests a passive reservation path in the local network. The foreign EA, which is in charge of that foreign mobile node, will inform the local EA of the SMR of the requesting foreign mobile node.

- If there are free standard channels in the local network, the local EA will allocate one to the foreign mobile node.
- If there is no free standard channel in the local network, the local EA will compare and find out whether the SMR of the requesting foreign mobile node is larger than the smallest of the SMRs of the foreign mobile nodes that have already been allocated standard channels. If it is, the foreign mobile node with the smallest SMR is replaced by the requesting foreign mobile node, i.e., the standard channel is re-allocated to the requesting mobile node. Otherwise, the passive reservation request is blocked.

Note that the SMR-based replacement mechanism should not be applied to active reservation requests because this would affect the ongoing QoS sessions. On the other hand, the re-allocation of passive reservation paths has no effect on the QoS sessions of the foreign mobile nodes since they are not actively using them.

5.3.3 Operation Procedure

In the following, the operation procedure of the reservation optimised ARR scheme from the perspective of the EA of the local network is described. Figure 5.1 shows the channel allocation procedure of the proposed scheme. The local EA receives two types of reservation requests.

- When the local EA receives an active reservation request from a local mobile node, it will allocate a free channel to that local mobile node (a dedicated channel is chosen first, or if one is not available, then a standard channel). When the local

[II]This can be achieved by receiving information messages from the mobile nodes, or by employing some statistical or history-based prediction algorithms. Typical examples of these can be found in [108–110], and they are beyond the scope of this work.

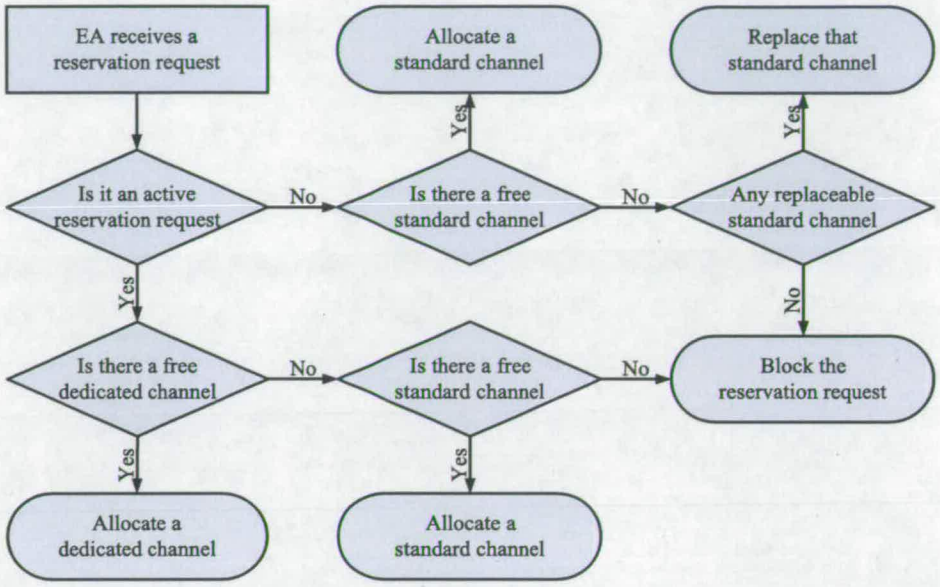


Figure 5.1: *The channel allocation procedure of the reservation optimised ARR scheme*

mobile node finishes its session, its active reservation path is released. Moreover, it also releases its active reservation path when it hands over out of the local network and instead it sends a passive reservation request to the local EA.

- When the local EA receives a passive reservation request from a foreign mobile node, it tries to allocate a standard channel to the foreign mobile node. The allocation procedure is described in Section 5.3.2. If the foreign mobile node fails to obtain a passive reservation path in the local network, it has to request an active reservation path when it hands over into the local network.

In brief, the reservation optimised ARR scheme is a CAC enhanced solution to the advance resource reservation problem. The CAC is carried out by the EA in each network by managing network resources with consideration of types of requests and mobility characteristics of mobile nodes. The motivation for integrating the CAC algorithm is to restrict the number of passive reservation paths in a network and only allow the most eligible mobile nodes to acquire them. In fact, to give an even higher priority to the active reservation requests, a passive reservation path can be replaced by an active reservation path. However, this kind of replacement is not included in the proposed scheme since it can result in an over-restriction on passive reservation paths in the network. In this work, a mobile node is considered to be more eligible if it has a larger SMR, with the assumption that all the sessions are of the same QoS class.

However, in a broader sense, different QoS classes should also be considered and it is a very important parameter to determine which mobile node is more suitable for being allocated a passive reservation path. This traffic class based admission control can be implemented in the *policy control* module of RSVP [24].

5.3.4 Modularity

To improve the efficiency of the reservation optimised ARR scheme, position prediction algorithms, which determine in which neighbouring networks a mobile node makes advance reservation paths, can be applied. With a precise position prediction algorithm and a low-cost signalling procedure such as RSVP aggregation [111], the signalling cost of the reservation optimised ARR scheme can be reduced.

Moreover, since the proposed scheme in fact consists of two admission control mechanisms, it can be easily integrated into existing ARR schemes by requiring them to implement these management mechanisms. As there are already agents in these ARR schemes, the only additional information required is the SMRs of the mobile nodes. Incidentally, collecting the session and mobility information of the mobile nodes is a basic requirement of next-generation wireless communications [112]. In this way, the modularity of the proposed scheme is maintained.

5.4 PEPA Models of the Conventional, Passive Reservation Limited and Reservation Optimised ARR Schemes

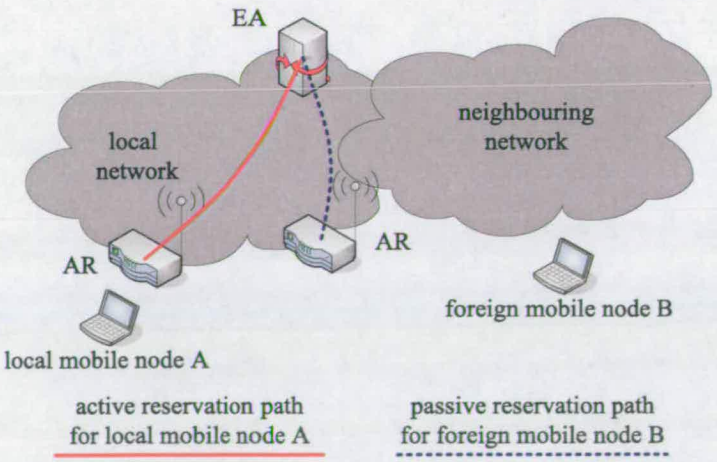


Figure 5.2: *Different types of reservation paths in the local network*

In this section, the PEPA models of the conventional and the reservation optimised ARR schemes are presented. Moreover, as a stepping stone between the two models, the PEPA model of the ARR scheme which only implements the passive reservation limited mechanism is also given. These models aim to express how the resources of a network are consumed by the mobile nodes. The network under observation is called a local network, and a mobile node is called a local and a foreign mobile node when it is in and out of the local network respectively. Figure 5.2 shows an example of the different types of reservation paths in the local network. The local mobile node A makes an active reservation path in the local network, whereas the foreign mobile node B makes a passive reservation path in the local network. Moreover, the assumptions about the traffic and mobility models of the mobile nodes are the same as those in Chapter 4.

5.4.1 PEPA Model of the Conventional ARR Scheme

In the conventional ARR scheme, a local mobile node requires an active reservation path in the local network, whereas a foreign mobile node requires a passive reservation path in the local network. The local network does not discriminate between the active and passive reservation requests, and the type of a reservation path is changed according to the movement of its reserver. There are two types of PEPA components in this model. The component MN models the behaviour of a mobile node, and a channel of the local network is modelled by the component $CHAN$.

Mobile Node: Since the resources of the local network are used by the local and foreign mobile nodes in different ways, the states of a mobile node are distinguished according to its position. Superscripts L and F are used to denote that the mobile node is in the local and neighbouring networks respectively. The mobile node is initially in the idle state MN^L_{Idle} (or state MN^F_{Idle}) and can move between different networks (activities $move_{outwards}$ and $move_{inwards}$). Its mean sojourn times in local and neighbouring networks are $1/\upsilon_{out}$ and $1/\upsilon_{in}$ respectively. The mobile node generates new session requests (activity $session\_arrive$) at the rate of λ. Depending on the position of the mobile node, i.e., in states $MN^L_{Request}$ and $MN^F_{Request}$, it requires an active or a passive reservation path in the local network respectively (activities $reserve_{active}$ and $reserve_{passive}$). In state $MN^L_{Request}$, if the mobile node's active reservation request is blocked, it keeps requesting until the resources become available, or during this time it may move out of the local network and change to state $MN^F_{Request}$. In state $MN^F_{Request}$,

if the mobile node's passive reservation request is blocked, it keeps requesting until its request is admitted, or its session is finished.[III] Note that in the latter case, the mobile node's session behaviour (activity *session*) has no effect on the local network. Alternatively, the mobile node may move into the local network and change to state $MN^L_{Request}$. At the engaged states $MN^L_{Engaged}$ and $MN^F_{Engaged}$, the mobile node actively and passively uses its reservation path respectively (activities $session_{active}$ and $session_{passive}$), with a mean duration of $1/\mu$. For the foreign mobile node which has successfully reserved resources in advance, it can continue its session when it hands over into the local network (activity $handover_{inwards}$). As for the local mobile node, when it hands over out of the local network (activity $handover_{outwards}$), it still occupies its reservation path in the local network, though in a passive way. When the mobile node finishes its session, it returns to the idle state. The component MN is defined as:

$$
\begin{aligned}
MN^L_{Idle} &\stackrel{def}{=} (session\_arrive, \lambda).MN^L_{Request} \\
&\quad + (move_{outwards}, v_{out}).MN^F_{Idle} \\
MN^L_{Request} &\stackrel{def}{=} (reserve_{active}, r_{active}).MN^L_{Engaged} \\
&\quad + (move_{outwards}, v_{out}).MN^F_{Request} \\
MN^L_{Engaged} &\stackrel{def}{=} (session_{active}, \mu).MN^L_{Idle} \\
&\quad + (handover_{outwards}, v_{out}).MN^F_{Engaged} \\
MN^F_{Idle} &\stackrel{def}{=} (session\_arrive, \lambda).MN^F_{Request} \\
&\quad + (move_{inwards}, v_{in}).MN^L_{Idle} \\
MN^F_{Request} &\stackrel{def}{=} (reserve_{passive}, r_{passive}).MN^F_{Engaged} \\
&\quad + (session, \mu).MN^F_{Idle} \\
&\quad + (move_{inwards}, v_{in}).MN^L_{Request} \\
MN^F_{Engaged} &\stackrel{def}{=} (session_{passive}, \mu).MN^F_{Idle} \\
&\quad + (handover_{inwards}, v_{in}).MN^L_{Engaged}
\end{aligned}
$$

Channel: A channel of the local network has three states: idle, active and passive. At state $CHAN_{Idle}$ the channel can accept active and passive reservation requests and go to the states $CHAN_{Active}$ and $CHAN_{Passive}$ respectively. The type of a reservation path is changed according to the movement of its reserver. When a mobile node finishes its session, its reservation path is released. The component $CHAN$ is defined

[III] In fact, a foreign mobile node cannot start its communication unless its active reservation request is admitted by its current network. To simplify the models, it is assumed that the foreign mobile node's active reservation requests are always admitted by its current network.

as:

$$\begin{aligned}
CHAN_{Idle} &\stackrel{def}{=} (reserve_{active}, \top).CHAN_{Active} \\
&\quad + (reserve_{passive}, \top).CHAN_{Passive} \\
CHAN_{Active} &\stackrel{def}{=} (session_{active}, \top).CHAN_{Idle} \\
&\quad + (handover_{outwards}, \top).CHAN_{Passive} \\
CHAN_{Passive} &\stackrel{def}{=} (session_{passive}, \top).CHAN_{Idle} \\
&\quad + (handover_{inwards}, \top).CHAN_{Active}
\end{aligned}$$

System Definition: The component $CHAN$ synchronises with the component MN on the reservation request activities $reserve_{active}$ and $reserve_{passive}$, and the resource holding activities $session_{active}$ and $session_{passive}$. Moreover, the type of a reservation path is changed according to the mobile node's handover activities $handover_{outwards}$ and $handover_{inwards}$. The PEPA model of the conventional ARR scheme which consists of K mobile nodes and X channels is constructed as:

$$ARR^{CON} \stackrel{def}{=} \left(MN^{L}_{Idle}[K]\right) \underset{L}{\bowtie} \left(CHAN_{Idle}[X]\right),$$

where

$$L = \left\{reserve_{active}, reserve_{passive}, session_{active}, session_{passive}, handover_{outwards}, handover_{inwards}\right\}.$$

5.4.2 PEPA Model of the Passive Reservation Limited ARR Scheme

In the passive reservation limited ARR scheme, dedicated channels are set aside especially for active reservation paths. Therefore, there are two PEPA components $CHAN^S$ and $CHAN^D$ to model the standard and dedicated channels respectively. The behaviour of the mobile node also needs to be modified to implement the passive reservation limited mechanism described in Section 5.3.1.

Mobile Node: Since a local mobile node can use both standard and dedicated channels, at state $MN^{L}_{Request}$ the local mobile node can either request a dedicated channel (activity $reserve_{prior}$), or request a standard channel when there is no free dedicated channel (activity $reserve_{active}$). Moreover, since the passive reservation limited scheme does not allow an active reservation path to change to the passive state, the engaged local mobile node has to request a passive reservation path when it hands over out of the local network, i.e., it goes to state $MN^{F}_{Request}$ instead of the state

$MN^F_{Engaged}$. The definitions of the other states and the behaviour of the component *MN* are the same as in the conventional ARR scheme model. The component *MN* is defined as:

$$
\begin{aligned}
MN^L_{Idle} &\stackrel{def}{=} (session\_arrive, \lambda).MN^L_{Request} \\
&\quad + (move_{outwards}, v_{out}).MN^F_{Idle} \\
MN^L_{Request} &\stackrel{def}{=} (reserve_{active}, r_{active}).MN^L_{Engaged} \\
&\quad + (reserve_{prior}, r_{prior}).MN^L_{Engaged} \\
&\quad + (move_{outwards}, v_{out}).MN^F_{Request} \\
MN^L_{Engaged} &\stackrel{def}{=} (session_{active}, \mu).MN^L_{Idle} \\
&\quad + (handover_{outwards}, v_{out}).MN^F_{Request} \\
MN^F_{Idle} &\stackrel{def}{=} (session\_arrive, \lambda).MN^F_{Request} \\
&\quad + (move_{inwards}, v_{in}).MN^L_{Idle} \\
MN^F_{Request} &\stackrel{def}{=} (reserve_{passive}, r_{passive}).MN^F_{Engaged} \\
&\quad + (session, \mu).MN^F_{Idle} \\
&\quad + (move_{inwards}, v_{in}).MN^L_{Request} \\
MN^F_{Engaged} &\stackrel{def}{=} (session_{passive}, \mu).MN^F_{Idle} \\
&\quad + (handover_{inwards}, v_{in}).MN^L_{Engaged}
\end{aligned}
$$

Standard Channel: An idle standard channel can accept a passive reservation request from a foreign mobile node (activity $reserve_{passive}$) and is passively reserved. When the foreign mobile node hands over into the local network, the passively reserved standard channel becomes active, i.e., goes to state $CHAN^S_{Active}$. On the other hand, an idle standard channel can also be actively reserved by a local mobile node when there is no free dedicated channel (activity $reserve_{active}$). However, an actively reserved standard channel is released, i.e., goes to state $CHAN^S_{Idle}$, when the local mobile node hands over out of the local network. An engaged standard channel is released when its reserver finishes its session. The component $CHAN^S$ is defined as:

$$
\begin{aligned}
CHAN^S_{Idle} &\stackrel{def}{=} (reserve_{active}, \top).CHAN^S_{Active} \\
&\quad + (reserve_{passive}, \top).CHAN^S_{Passive} \\
CHAN^S_{Active} &\stackrel{def}{=} (session_{active}, \top).CHAN^S_{Idle} \\
&\quad + (handover_{outwards}, \top).CHAN^S_{Idle} \\
CHAN^S_{Passive} &\stackrel{def}{=} (session_{passive}, \top).CHAN^S_{Idle} \\
&\quad + (handover_{inwards}, \top).CHAN^S_{Active}
\end{aligned}
$$

Dedicated Channel: An idle dedicated channel can only accept active reservation requests from a local mobile node (activity $reserve_{prior}$) and be actively reserved. An engaged dedicated channel is released when its reserver finishes its session or hands over out of the local network. To guarantee that the local mobile node chooses the dedicated channels before the standard channels, the activity $reserve_{active}$ is defined as a self-transition at state $CHAN^D_{Active}$ and all the components $CHAN^D$ are required to cooperate on it. In this way, reserving a standard channel (activity $reserve_{active}$) is only enabled when all the dedicated channels are engaged. The component $CHAN^D$ is defined as:

$$\begin{aligned} CHAN^D_{Idle} &\stackrel{def}{=} (reserve_{prior}, \top).CHAN^D_{Active} \\ CHAN^D_{Active} &\stackrel{def}{=} (session_{active}, \top).CHAN^D_{Idle} \\ &\quad + (handover_{outwards}, \top).CHAN^D_{Idle} \\ &\quad + (reserve_{active}, \top).CHAN^D_{Active} \end{aligned}$$

System Definition: As in the conventional ARR scheme model, the component MN synchronises with the components $CHAN^S$ and $CHAN^D$ on the reservation request, resource holding and handover activities. Moreover, all the components $CHAN^D$ synchronise with each other and with the components $CHAN^S$ on the activity $reserve_{active}$, which guarantees that the dedicated channels are selected first. The PEPA model of the passive reservation limited ARR scheme which consists of K mobile nodes, X standard channels and Y dedicated channels is constructed as:

$$ARR^{PRL} \stackrel{def}{=} \left(MN^L_{Idle}[K]\right) \underset{L_1}{\bowtie} \left(\left(CHAN^S_{Idle}[X]\right) \underset{L_2}{\bowtie} \left(\underbrace{CHAN^D_{Idle} \underset{L_2}{\bowtie} CHAN^D_{Idle} \cdots \underset{L_2}{\bowtie} CHAN^D_{Idle}}_{Y}\right)\right),$$

where

$$L_1 = \left\{reserve_{active}, reserve_{prior}, reserve_{passive}, session_{active}, session_{passive}, handover_{outwards}, handover_{inwards}\right\}, \text{ and } L_2 = \left\{reserve_{active}\right\}.$$

5.4.3 PEPA Model of the Reservation Optimised ARR Scheme

The reservation optimised ARR scheme includes the passive reservation limited and the SMR-based replacement mechanisms. The objective of the SMR-based replacement mechanism is to make the best usage of the standard channels in a network. Only the foreign mobile nodes with the highest SMRs are eligible to make passive reservation paths in the local network. Since the SMR of a mobile node is dynamic and the replacement procedure is hard to implement using the performance modelling formalisms, an approach that *approximate* the replacement procedure is employed. That is, the SMR-based replacement mechanism forbids some mobile nodes from requesting passive reservations. Therefore, the mobile nodes in the network are classified into two groups: the *fast* mobile nodes and the *slow* mobile nodes. This type of classification does not lose generality since there will always be some mobile nodes that have higher SMRs than the others and are eligible to request passive reservation paths. These mobile nodes can be regarded as the fast mobile nodes and the rest can be regarded as the slow mobile nodes. The fast and slow mobile nodes are modelled by the components *Fast_MN* and *Slow_MN* respectively.

Fast Mobile Node: A fast mobile node can make both active and passive reservation requests. Its behaviour is the same as the component *MN* in the passive reservation limited ARR scheme model. The component *Fast_MN* is defined as:

$$
\begin{aligned}
Fast\_MN^{L}_{Idle} &\stackrel{def}{=} (session\_arrive, \lambda).Fast\_MN^{L}_{Request} \\
&\quad + (move_{outwards}, v_{out}).Fast\_MN^{F}_{Idle} \\
Fast\_MN^{L}_{Request} &\stackrel{def}{=} (reserve_{active}, r_{active}).Fast\_MN^{L}_{Engaged} \\
&\quad + (reserve_{prior}, r_{prior}).Fast\_MN^{L}_{Engaged} \\
&\quad + (move_{outwards}, v_{out}).Fast\_MN^{F}_{Request} \\
Fast\_MN^{L}_{Engaged} &\stackrel{def}{=} (session_{active}, \mu).Fast\_MN^{L}_{Idle} \\
&\quad + (handover_{outwards}, v_{out}).Fast\_MN^{F}_{Request} \\
Fast\_MN^{F}_{Idle} &\stackrel{def}{=} (session\_arrive, \lambda).Fast\_MN^{F}_{Request} \\
&\quad + (move_{inwards}, v_{in}).Fast\_MN^{L}_{Idle} \\
Fast\_MN^{F}_{Request} &\stackrel{def}{=} (reserve_{passive}, r_{passive}).Fast\_MN^{F}_{Engaged} \\
&\quad + (session, \mu).Fast\_MN^{F}_{Idle} \\
&\quad + (move_{inwards}, v_{in}).Fast\_MN^{L}_{Request} \\
Fast\_MN^{F}_{Engaged} &\stackrel{def}{=} (session_{passive}, \mu).Fast\_MN^{F}_{Idle} \\
&\quad + (handover_{inwards}, v_{in}).Fast\_MN^{L}_{Engaged}
\end{aligned}
$$

Slow Mobile Node: A slow mobile node behaves differently from a fast mobile node when it is in the neighbouring networks. The passive reservation requests of the slow foreign mobile node are always blocked by the local network (state $Slow\_MN^F_{Blocked}$), and it stops requesting until it finishes its session[IV] or moves into the local network and requests an active reservation path. Moreover, since the slow mobile node in the neighbouring networks has no effect on the resource utilisation of the local network, at state $Slow\_MN^F_{Blocked}$ its session and handover behaviour are named differently (activities *session* and *handover*) so that they do not synchronise with the channels of the local network. Note that if the slow mobile node hands over into the local network before the end of its session, it needs to request an active reservation path since it does not have resources reserved in advance. The component *Slow_MN* is defined as:

$$
\begin{aligned}
Slow\_MN^L_{Idle} &\stackrel{def}{=} (session\_arrive, \lambda).Slow\_MN^L_{Request} \\
&\quad + (move_{outwards}, v^{slow}_{out}).Slow\_MN^F_{Idle} \\
Slow\_MN^L_{Request} &\stackrel{def}{=} (reserve_{active}, r_{active}).Slow\_MN^L_{Engaged} \\
&\quad + (reserve_{prior}, r_{prior}).Slow\_MN^L_{Engaged} \\
&\quad + (move_{outwards}, v^{slow}_{out}).Slow\_MN^F_{Blocked} \\
Slow\_MN^L_{Engaged} &\stackrel{def}{=} (session_{active}, \mu).Slow\_MN^L_{Idle} \\
&\quad + (handover_{outwards}, v^{slow}_{out}).Slow\_MN^F_{Blocked} \\
Slow\_MN^F_{Idle} &\stackrel{def}{=} (session\_arrive, \lambda).Slow\_MN^F_{Blocked} \\
&\quad + (move_{inwards}, v^{slow}_{in}).Slow\_MN^L_{Idle} \\
Slow\_MN^F_{Blocked} &\stackrel{def}{=} (session, \mu).Slow\_MN^F_{Idle} \\
&\quad + (handover, v^{slow}_{in}).Slow\_MN^L_{Request}
\end{aligned}
$$

Standard Channel and Dedicated Channel: The definitions of the components $CHAN^S$ and $CHAN^D$ are the same as those in the passive reservation limited ARR scheme model and they are omitted here.

System Definition: The cooperation relations between the mobile nodes and the channels in this model are the same as those in the passive reservation limited ARR scheme model. Moreover, the number of standard channels is the maximum number of mobile nodes that are eligible to make passive reservation paths, and thus is also the number of the mobile nodes that can be regarded as the fast mobile nodes. Therefore,

[IV] See footnote III.

the PEPA model of the reservation optimised ARR scheme which consists of X fast mobile nodes, Z slow mobile nodes, X standard channels and Y dedicated channels is constructed as:

$$ARR^{RO} \stackrel{def}{=} \Big(\Big(Fast\_MN^{L}_{Idle}[X]\Big) \| \Big(Slow\_MN^{L}_{Idle}[Z]\Big)\Big) \underset{L_1}{\bowtie} \Big(\Big(\bar{CHAN}^{S}_{Idle}[\bar{X}]\Big) \underset{L_2}{\bowtie} \Big(\underbrace{CHAN^{D}_{Idle} \underset{L_2}{\bowtie} CHAN^{D}_{Idle} \cdots \underset{L_2}{\bowtie} CHAN^{D}_{Idle}}_{Y}\Big)\Big),$$

where

$$L_1 = \Big\{reserve_{active}, reserve_{prior}, reserve_{passive}, session_{active}, session_{passive}, handover_{outwards}, handover_{inwards}\Big\}, \text{ and } L_2 = \Big\{reserve_{active}\Big\}.$$

5.4.4 System States of the PEPA Models

Like in Chapter 4, to guarantee the models are numerically tractable, all the models have 4 mobile nodes and 3 channels. In the passive reservation limited ARR scheme model there are 2 standard channels and 1 dedicated channel. In the reservation optimised ARR scheme model there are 2 fast mobile nodes, 2 slow mobile nodes, 2 standard channels and 1 dedicated channel. The system states of all the models are the *feasible* combinations of the state of each component of that model. Like in Chapter 4, AWK codes are used to check the feasibility of system states of all the models. For instance, the number of engaged local mobile nodes should equal to the number of active reservation paths; the number of engaged (fast) foreign mobile nodes and that of passive reservation paths should be the same; and dedicated channels should be allocated before standard channels for active reservations.

As examples, the following 3 system states are selected from the models respec-

tively:

For the conventional ARR scheme model:

$$\left(MN^{F}_{Engaged} | MN^{F}_{Request} | MN^{L}_{Engaged} | MN^{L}_{Idle} | CHAN_{Passive} | CHAN_{Active} | CHAN_{Idle}\right);$$

for the passive reservation limited ARR scheme model:

$$\left(MN^{F}_{Idle} | MN^{F}_{Engaged} | MN^{L}_{Request} | MN^{L}_{Idle} | CHAN^{S}_{Passive} | CHAN^{S}_{Idle} | CHAN^{D}_{Idle}\right);$$

and for the reservation optimised ARR scheme model:

$$\left(Fast\_MN^{F}_{Idle} | Fast\_MN^{F}_{Engaged} | Slow\_MN^{L}_{Idle} | Slow\_MN^{F}_{Idle} | CHAN^{S}_{Passive} | CHAN^{S}_{Idle} | CHAN^{D}_{Idle}\right).$$

5.5 Performance Evaluation

The fundamental goal of this work is to investigate how the resources of a network can be managed when they are under-provisioned. The effect of traffic intensity on the blocking probabilities of active and passive reservation requests are evaluated. These performance measures are of interest because they reflect the network congestion levels for different types of reservation paths. Moreover, the mean numbers of active and passive reservation paths in a network are also evaluated to investigate the effect of different ARR schemes on the resource utilisation of a network.

5.5.1 Parameter Settings

The activity rates used in the PEPA models of the conventional, passive reservation limited, and reservation optimised ARR schemes are set as follows. The traffic intensity is tuned by the session arrival interval $1/\lambda$ and the session duration $1/\mu$, and their variation ranges are listed in Table 5.1. The mean residence times of a (fast) mobile node within and outside the local network are set to 480 and 960 seconds respectively.[v] As for a slow mobile node, its sojourn time in an area is twice as long as that of its fast counterpart. The mean delay of the active reservation request messages is set to 0.1s, and since the passive reservation requests are sent from the neighbouring networks, their mean delay is set to 0.2s. The rates of all the activities are listed in Table 5.1.

[v] This is a contrived assumption that the time spent by the mobile node staying in the local network is half of the time it spends staying outside of the local network.

Rate	Corresponding activities	Value (second^{-1})
λ	$session\_arrive$	$[30:30:450].^{-1}$
μ	$session_{active}$, $session_{passive}$, $session$	$[45:45:675].^{-1}$
υ_{out}	$move_{outwards}$, $handover_{outwards}$	480^{-1}
υ_{out}^{slow}	$move_{outwards}$, $handover_{outwards}$	960^{-1}
υ_{in}	$move_{inwards}$, $handover_{inwards}$	960^{-1}
υ_{in}^{slow}	$move_{inwards}$, $handover_{inwards}$, $handover$	1920^{-1}
r_{active}	$reserve_{active}$	0.1^{-1}
r_{prior}	$reserve_{prior}$	0.1^{-1}
$r_{passive}$	$reserve_{passive}$	0.2^{-1}

Table 5.1: *Activity rates of the PEPA models of the conventional, passive reservation limited, and reservation optimised ARR schemes*

5.5.2 Active Reservation Blocking Probability

The first *MN* component in the conventional and passive reservation limited ARR scheme models, and the first *Fast_MN* and *Slow_MN* in the reservation optimised ARR scheme model are chosen to be investigated. The active reservation blocking happens in the system states in which the mobile node is in the local network and requires an active reservation path whilst no free channel is available. Therefore, in the conventional ARR scheme model, these system states can be described as the set of system states:

$$S_{ARB}^{CON} = \left\{ \left(MN_{Request}^{L} | * | * | * | ! CHAN_{Idle} | ! CHAN_{Idle} | ! CHAN_{Idle} \right) \right\},$$

and in the passive reservation limited ARR scheme model, these system states can be described as the set of system states:

$$S_{ARB}^{PRL} = \left\{ \left(MN_{Request}^{L} | * | * | * | ! CHAN_{Idle}^{S} | ! CHAN_{Idle}^{S} | ! CHAN_{Idle}^{D} \right) \right\},$$

and for the fast and slow mobile nodes in the reservation optimised ARR scheme model, these system states can be described as the set of system states:

$$S_{ARB,Fast}^{RO} = \left\{ \left(Fast\_MN_{Request}^{L} | * | * | * | ! CHAN_{Idle}^{S} | ! CHAN_{Idle}^{S} | ! CHAN_{Idle}^{D} \right) \right\}, \text{ and}$$

$$S^{RO}_{ARB,Slow} = \left\{ \left(* \mid * \mid Slow\_MN^{L}_{Request} \mid * \mid !CHAN^{S}_{Idle} \mid !CHAN^{S}_{Idle} \mid !CHAN^{D}_{Idle} \right) \right\}$$

respectively. The symbol "!" is used as a "not" character and means that the component can be in any state other than the one after the "!" symbol. In these system states, the mobile node waits for the resources to become available before it moves out of the local network. Hence, the percentage of time that the mobile node spends on waiting, which can also be regarded as the active reservation blocking probability experienced by the mobile node, in each model is calculated as:

$$P^{CON}_{ARB} = \sum_{s_i \in S^{CON}_{ARB}} \pi(s_i), \tag{5.1}$$

$$P^{PRL}_{ARB} = \sum_{s_i \in S^{PRL}_{ARB}} \pi(s_i), \tag{5.2}$$

$$P^{RO}_{ARB,Fast} = \sum_{s_i \in S^{RO}_{ARB,Fast}} \pi(s_i), \tag{5.3}$$

$$P^{RO}_{ARB,Slow} = \sum_{s_i \in S^{RO}_{ARB,Slow}} \pi(s_i), \tag{5.4}$$

respectively, where $\pi(s_i)$ is the equilibrium probability of the system state s_i.

5.5.2.1 Effect of Session Arrival Interval

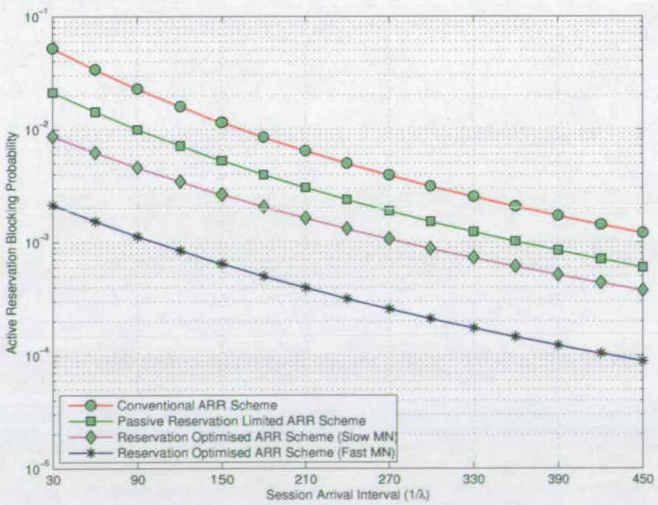


Figure 5.3: *The effect of session arrival interval on active reservation blocking probability*

Figure 5.3 shows the effect of session arrival interval on the active reservation blocking probability for the three schemes. The results are calculated with the mean

session duration set to 360 seconds, and the y-axis is in the logarithmic scale. It can be observed from the figure that the active reservation blocking probability decreases when the session arrives less frequently. The passive reservation limited ARR scheme has a lower blocking probability (from 2.12×10^{-2} to 6.00×10^{-4}) than the conventional ARR scheme (from 5.23×10^{-2} to 1.21×10^{-3}), because it sets aside dedicated channels for active reservation requests. In the reservation optimised ARR scheme, since fewer foreign mobile nodes are eligible to request standard channels for making passive reservation paths, the competition for the resources in the local network is less severe. Therefore, both fast and slow mobile nodes in the reservation optimised ARR scheme have a lower active reservation blocking probability than the other two ARR schemes. Moreover, as a slow mobile node stays longer in the local network than a fast mobile node, at the same session arrival interval, it generates more requests during its sojourn time in the local network. For that reason, the slow mobile node is more likely to be rejected (from 8.69×10^{-3} to 3.76×10^{-4}) than the fast mobile node (from 2.14×10^{-3} to 8.90×10^{-5}) for active reservation requests.

5.5.2.2 Effect of Session Duration

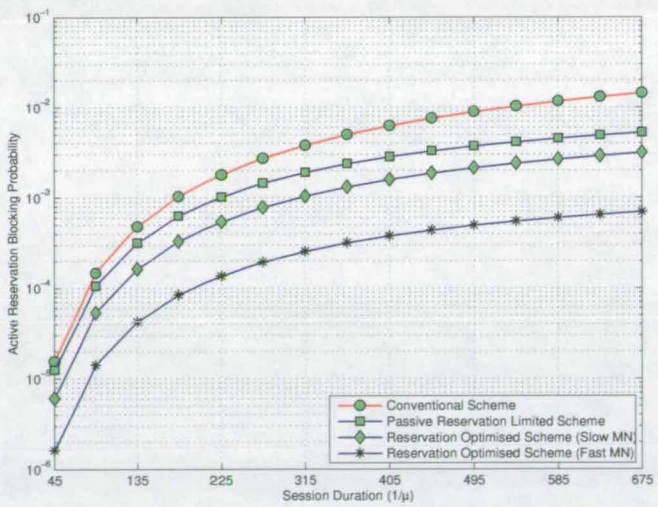


Figure 5.4: *The effect of session duration on active reservation blocking probability*

A similar improvement on the active reservation blocking probability in the passive reservation limited and reservation optimised schemes can also be observed in Figure 5.4, where the session arrival interval is set to 240 seconds and the session duration is changed. It is clear that the engaged mobile nodes hold the resources

for a longer time when the session duration grows, and thus the active reservation blocking probability increases. When the session duration is less than 90 seconds, the passive reservation limited ARR scheme and the conventional ARR scheme have close performance on the active reservation blocking. However, as the session duration gets larger, the former grows from 1.25×10^{-5} to 5.28×10^{-3} and clearly outperforms the latter which increases from 1.56×10^{-5} to 1.45×10^{-2}. Again, the reservation optimised ARR scheme has the lowest active reservation blocking probability, in which the slow mobile node has a blocking probability growing from 6.09×10^{-6} to 3.17×10^{-3} and the fast mobile node has a lower blocking probability ranging from 1.63×10^{-6} to 7.00×10^{-4}.

5.5.3 Passive Reservation Blocking Probability

The passive reservation blocking happens in the system states in which the mobile node is in the neighbouring networks and requires a passive reservation path in the local network but no free channel is available. In the conventional ARR scheme model, all the channels can be used for passive reservation paths, and thus these system states can be described as the set of system states:

$$S_{PRB}^{CON} = \left\{ \left(MN_{Request}^{F} | * | * | * | !CHAN_{Idle} | !CHAN_{Idle} | !CHAN_{Idle} \right) \right\},$$

and in the passive reservation limited ARR scheme model, since only standard channels can be used for passive reservation paths, these system states can be described as the set of system states:

$$S_{PRB}^{PRL} = \left\{ \left(MN_{Request}^{F} | * | * | * | !CHAN_{Idle}^{S} | !CHAN_{Idle}^{S} | * \right) \right\},$$

and similarly in the reservation optimised ARR scheme model, the fast mobile node is unable to make passive reservation paths in the set of system states:

$$S_{PRB,Fast}^{RO} = \left\{ \left(Fast\_MN_{Request}^{F} | * | * | * | !CHAN_{Idle}^{S} | !CHAN_{Idle}^{S} | * \right) \right\},$$

and for the slow mobile node, since it is always blocked when it is out of the local network, its passive reservation blocking probability is not investigated. In each model, the passive reservation blocking probability experienced by the mobile node

is calculated as:

$$P_{PRB}^{CON} = \sum_{s_i \in S_{PRB}^{CON}} \pi(s_i), \tag{5.5}$$

$$P_{PRB}^{PRL} = \sum_{s_i \in S_{PRB}^{PRL}} \pi(s_i), \tag{5.6}$$

$$P_{PRB,Fast}^{RO} = \sum_{s_i \in S_{PRB,Fast}^{RO}} \pi(s_i), \tag{5.7}$$

respectively. As in Section 5.5.2, the effect of the traffic intensity on the passive reservation blocking probability is investigated, and the activity rates for each of the following subsections are the same as their counterpart in Section 5.5.2.

5.5.3.1 Effect of Session Arrival Interval

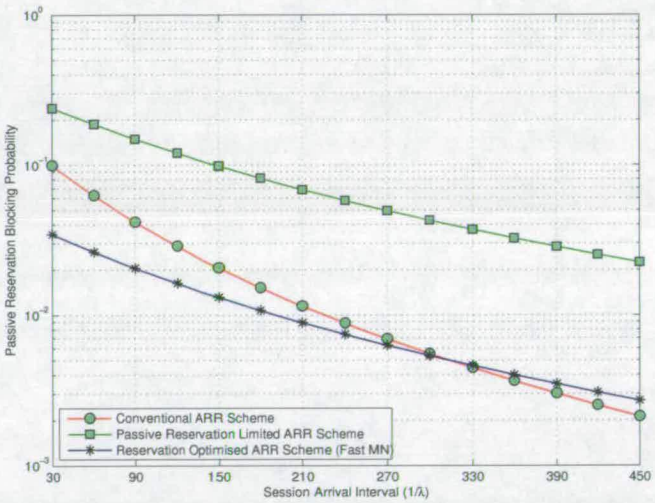


Figure 5.5: *The effect of session arrival interval on passive reservation blocking probability*

Figs. 5.5 shows how passive reservation requests of the mobile nodes are affected by the restrictions on the passive reservations. The conventional ARR scheme has a passive reservation blocking probability ranging from 9.97×10^{-2} to 2.12×10^{-3}. Since the passive reservation limited ARR scheme restricts the resource for passive reservations, it has a higher blocking probability than the conventional ARR scheme which decreases from 2.40×10^{-1} to 2.25×10^{-2}. An interesting observation is that the fast mobile node in the reservation optimised ARR scheme has a passive reservation blocking probability that ranges from 3.46×10^{-2} to 2.70×10^{-3} and is smaller than that

in the conventional ARR scheme when the session arrival interval is small. Remember that in the reservation optimised ARR scheme, the resource competition is less severe since not all the foreign mobile nodes are allowed to make passive reservation paths. The results indicate that at high traffic intensities, the resource competition caused by the number of the mobile nodes has greater effect than that caused by limited resource. In other words, reducing the number of requests can compensate the resource restriction. However, this difference decreases and at session arrival intervals larger than 330 seconds, the fast mobile node experiences larger blocking probability in the reservation optimised ARR scheme, which indicates that the limited resource now has larger effect on the passive reservation blocking probability.

5.5.3.2 Effect of Session Duration

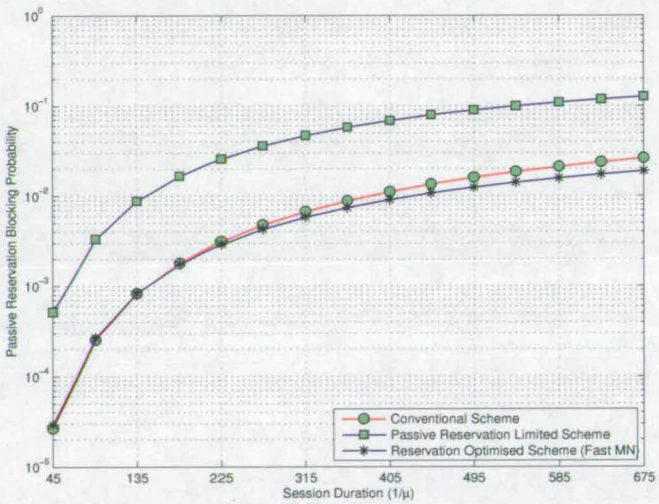


Figure 5.6: *The effect of session duration on passive reservation blocking probability*

Similar results can be observed in Figure 5.6 where the session duration is the control parameter. The passive reservation limited ARR scheme has the highest passive reservation blocking probability that increases from 5.20×10^{-4} to 1.27×10^{-1}. The fast mobile node in the reservation optimised ARR scheme has a passive reservation blocking probability ranging from 2.91×10^{-5} to 1.90×10^{-2}, which is very close to that in the conventional ARR scheme ranging from 2.67×10^{-5} to 2.64×10^{-2}. However, the former outperforms the latter at higher traffic intensities, i.e., when the session duration is larger than 225 seconds.

5.5.4 Mean Numbers of Active and Passive Reservation Paths

The mean numbers of the active and passive reservation paths in a network are also of interest, as they reflect how different ARR schemes affect the resource utilisation of a network. Using MRS and Eq. (3.3), the mean number of a certain type of reservation paths can be found by setting the reward ρ_i equal to the number of that type of reservation paths in the system state s_i. That is, the mean number of active reservation paths is calculated as:

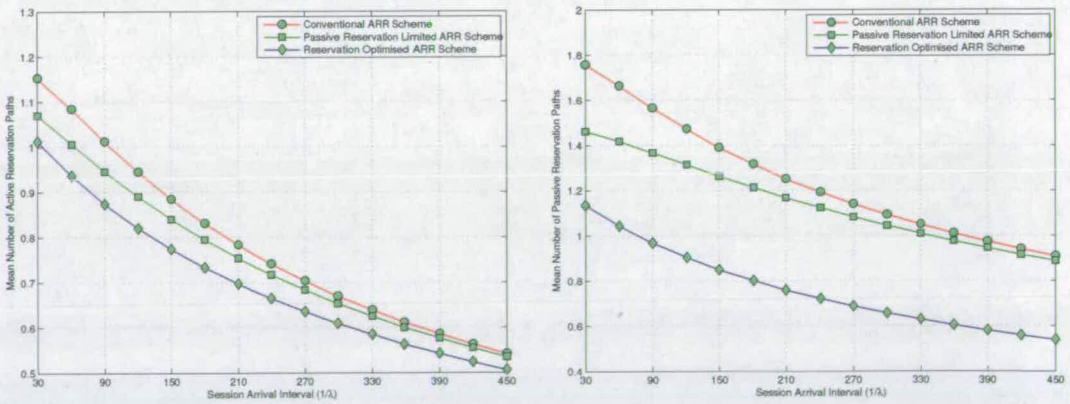
$$N^{Active} = \sum_{s_i \in S} n_i^{Active} * \pi(s_i), \tag{5.8}$$

where n_i^{Active} is the number of active reservation paths in the system state s_i, and S is the whole system state space of a model. Similarly, the mean number of passive reservation paths is calculated as:

$$N^{Passive} = \sum_{s_i \in S} n_i^{Passive} * \pi(s_i). \tag{5.9}$$

The effect of the traffic intensity on the resource utilisation is investigated, and the activity rates are the same as in previous subsections.

5.5.4.1 Effect of Session Arrival Interval



(a) Mean number of active reservation paths (b) Mean number of passive reservation paths

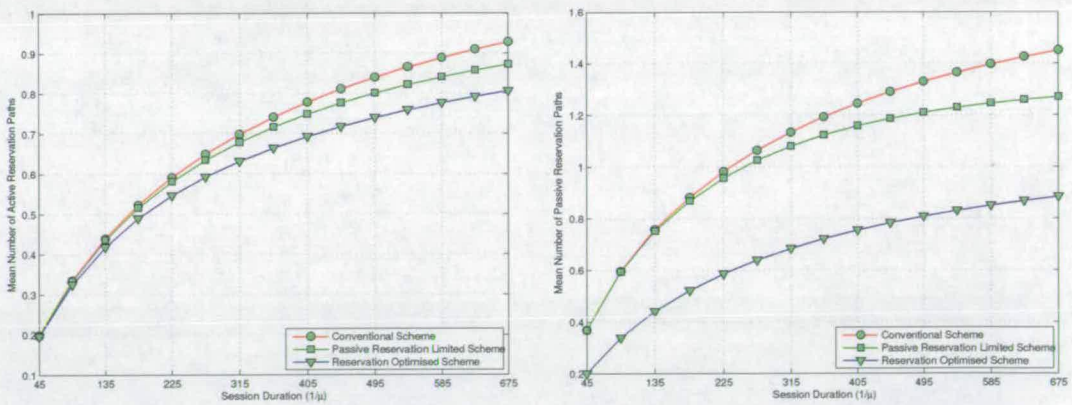
Figure 5.7: *The effect of session arrval interval on mean numbers of active and passive reservation paths*

Figure 5.7(a) shows the effect of session arrival interval on the mean number of ac-

tive reservation paths in a network. An interesting observation is that the conventional ARR scheme has the largest number of active reservation paths in a network, whereas the reservation optimised ARR scheme ranks last. This indicates that although the passive reservation limited ARR scheme and the reservation optimised ARR scheme give a higher priority to the active reservation requests, this does not result in a higher proportion of active reservation paths in the network. In fact, at the same traffic intensity, since all types of channels can be used for active reservation paths, the active reservation blocking probability can reflect the traffic load of the network, and the results indicate that the mean number of active reservation paths may be dependent on this blocking probability.

Similar trends can be seen in Figure 5.7(b), the passive reservation limited ARR scheme has a smaller number of passive reservation paths than the conventional ARR scheme because of the restriction on the available resources for passive reservation paths. The reservation optimised ARR scheme reduces this number further by preventing slow mobile nodes from making passive reservation paths.

5.5.4.2 Effect of Session Duration



(a) Mean number of active reservation paths (b) Mean number of passive reservation paths

Figure 5.8: *The effect of session duration on mean numbers of active and passive reservation paths*

Figure 5.8(a) and Figure 5.8(b) show the relationship between the mean numbers of active and passive reservation paths in a network and the session duration. Again, the conventional ARR scheme has the largest numbers of both active and passive reservation paths, followed by the passive reservation limited and then the reservation

optimised ARR schemes. However, the results of the conventional and the passive reservation limited ARR schemes are almost the same at session durations less than 135 seconds.

5.6 Conclusions

In addition to the signalling optimisation problem, another major problem of deploying RSVP in mobile wireless networks is the advance resource reservation problem. In this chapter, a novel reservation optimised ARR scheme which aims to balance the number of active and passive reservation paths in a network is proposed. The motivation is that the passive reservation paths in a network are not actively used by their owners and thus they waste network resources from the perspective of the QoS traffic. To demonstrate the advantages of the proposed ARR scheme, the performance of different ARR schemes have been compared from the perspectives of the active and passive reservation blocking probabilities, and the mean numbers of active and passive reservation paths in a network. These performance measures are investigated because they reflect the network congestion levels for different types of reservation paths and the resource utilisation in a network.

The results indicate that the proposed reservation optimised ARR scheme improves active reservation blocking probability and balances the active and passive reservation blocking probabilities effectively. This is attained by setting aside dedicated channels for active reservation paths and restricting the ability of some foreign mobile nodes to make passive reservation paths. Although the performance improvement is gained at the expense of introducing possible handover interruptions to the slow mobile nodes, the proposed scheme is still reasonable since:

1. Passive reservation blocking only means that a foreign mobile node cannot make an advance reservation path in the local network. Although there could be an interruption when this foreign mobile node hands over into the local network, this type of reservation blocking has no effect on its current QoS session. On the other hand, an active reservation request implies that there is a local mobile node which really needs the requested resources to start its communication. Therefore, it is practical to give a higher priority to the active reservation requests.

2. When a foreign mobile node that is not granted a passive reservation path hands over into the local network, it becomes a local mobile node and requires an active

reservation path. Since the proposed scheme sets aside dedicated channels for active reservation paths and reduces the active reservation blocking probability dramatically, this new local mobile node is more likely to acquire an active reservation path and continue its communication.

Together with the signalling optimisation schemes, the ARR schemes can improve the performance of the basic RSVP when it is deployed in mobile wireless networks. Certainly, carefully designed signalling procedures are necessary to guarantee the efficient integration of mobility and QoS in terms of signalling complexity and overheads. In the work presented in Chapter 4 and Chapter 5, there is no assumption regarding the access technologies used in the mobile wireless networks and they are transparent to the studied network management schemes. However, as presented in Chapter 2, the main trend for future wireless communications is a shift to heterogeneous wireless network environments in which handover management between different types of networks is a key to seamless mobility. In the next chapter, a general performance evaluation framework for strategies that are used by mobile nodes to select networks in this environment will be studied.

Chapter 6
Modelling of Network Selection Strategies in 3G and WLAN Interworking Networks

The work in the previous two chapters focuses on how mobility and QoS management protocols can be integrated to provide seamless services to mobile nodes in mobile wireless networks, without consideration of network access technologies. In this chapter, the diversity of access technology in a particularly popular heterogeneous wireless network environment, i.e. 3G-WLAN interworking networks, is taken into account, and a general performance evaluation framework for NSSs in this environment is investigated.

6.1 Introduction

With the rapid development of various wireless communication technologies, the main trend for future wireless communications is a shift from voice and text based services provided by early cellular networks to multimedia-based services provided by multiple and heterogeneous wireless networks. These multimedia applications require different bearer services and there is no single technology that can simultaneously provide low latency, cheap cost, large data rate and high mobility to mobile users [113]. Therefore, next-generation wireless communications focus on the integration of existing cellular and other wireless networks. Especially, owing to the recent evolution and successful deployment of the WLANs, there has been a demand for integrating WLANs with 3G WWANs, i.e. 3G-WLAN interworking. For example, the 3rd generation partnership project (3GPP), an international organisation which produces technical specifications and reports for 3G WWANs, has developed an architecture for the integration of the 3GPP cellular networks and WLANs with the intention to extend 3GPP services to the WLAN access environment [114].

In a 3G-WLAN interworking network, a mobile node may perform vertical handovers (VHOs) when it moves across overlaid 3G and WLAN radio access

networks (RANs). During the handover, the mobile node may use its NSS to select a certain type of radio access technology (RAT) to continue its communication. Since the NSS controls the session behaviour of the mobile node, it is important and meaningful to investigate how it affects the performance of user applications at the mobile node and the utilisation of the RANs. The contribution of the work presented in this chapter is that a general performance evaluation framework for the NSSs in 3G-WLAN interworking networks is constructed. This framework captures the traffic and mobility characteristics of the mobile nodes and has a good expression of the behaviour of the mobile nodes using different selection strategies. From the framework, important performance measures including average throughput, RAN blocking probability and handover rate are derived.

The rest of this chapter is structured as follows. Section 6.2 gives a brief review of the previous work on evaluating NSSs used in 3G-WLAN interworking networks. A traffic model of multimedia communications and a mobility model suitable for 3G-WLAN interworking networks which are used in the framework are presented in Section 6.3 and Section 6.4 respectively. In Section 6.5, PEPA models of the mobile nodes using different NSSs are presented. In Section 6.6, an iterative method that interlinks the PEPA models and a network resource consumption model in order to derive the RAN blocking probability and handover rate is described. The performance of different NSSs are evaluated in Section 6.7 and in Section 6.8 the evaluation results are discussed.

6.2 Related Work

In previous studies on the design and evaluation of NSSs in 3G-WLAN interworking networks, the selection strategies are first formulated as policy functions which take network and mobile node conditions as the input parameters and then output the preferred RAN. Then comparisons are made between the results of different strategies, or between the results of the same strategy under different scenarios, in terms of throughput, connection cost, power consumption, etc. Typical examples of this can be found in [41,46,115]. There are also studies which aim to integrate resource management schemes into the NSSs. In [116], network selection takes account of the traffic loads of different RANs in order to achieve a balance between the RANs. In [117, 118], the mobility, traffic and location information of the mobile nodes are

considered and network selection is implemented in a centralised way to manage the resources of the RANs more efficiently. An analytical model for evaluating different NSSs is investigated in [119]. Each strategy is formulated as a mapping function between the total session arrival rates of the whole 3G-WLAN interworking network and the session arrival rates into different RANs. However, this work only considers network selection for new session requests and node mobility is not taken into account.

One major limitation of the above studies is that the evaluation is usually carried out in a restricted way. That is, different strategies are compared at the policy function level without considering the mobile node's session and handover behaviour, whereas evaluation with consideration of traffic and mobility only focuses on a certain type of strategy. However, to make a fair comparison, different strategies should be evaluated using the same general framework with as few restrictions as possible. Therefore, the performance evaluation framework built in this work has the following characteristics:

- This framework has a good expression of the behaviour of a mobile node using different NSSs. To make the description more accurate, a traffic model which represents features of multimedia-based services, and a mobility model which captures movement characteristics of a mobile node within and across the 3G-WLAN interworking cells, are employed.

- NSSs for both new and handover sessions are considered in the framework, and they are embedded in the form of network selection probabilities. This level of abstraction provides the framework the flexibility to express any type of NSS given that its probabilities of choosing RANs can be determined.

- The generality of the framework is also retained by having an interface to the model capturing RAN resource consumption, in the form of the network blocking probability. In this way, the framework is independent on the CAC and resource management schemes used in the 3G-WLAN interworking networks.

6.3 Traffic Model

As in the previous two chapters, the traffic model is built at the session level and includes two parameters: session arrival rate and session duration. Owing to the multi-service characteristic of next-generation wireless communications, a mobile node may require different types of services with different durations. This

combination of various services results in a large variation in the observed user session duration. Moreover, the field data suggests that the statistical session duration in the Internet has a coefficient of variation (CoV) larger than one [120, 121]. To capture this traffic feature, the hyper-exponential distribution is employed to model the session duration.

A K-phase ($K > 1$) hyper-exponential distribution is composed of K exponentially distributed phases in parallel, and always has a CoV larger than one. To simplify the traffic model, a two-phase hyper-exponential distribution is used, where one phase represents the non-real time (NRT) sessions and the other represents the real-time (RT) sessions. The NRT sessions generally include Web surfing and file transfer, and the RT sessions generally include voice and video streaming. As for the session arrival rate, the general consensus that the session arrival is a Poisson process is followed.

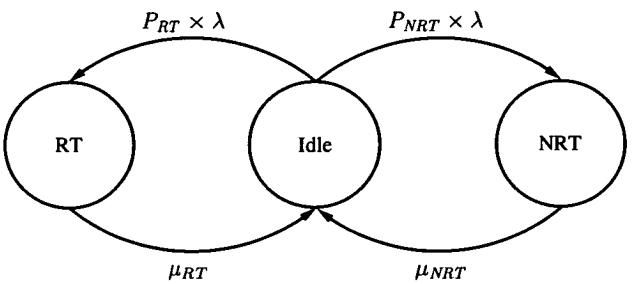


Figure 6.1: *A traffic model with two ON-OFF sources*

On the basis of the above assumptions, the traffic model can be constructed as a combination of two ON-OFF sources. As shown in Fig 6.1, the session arrival rate is λ and the session can be either an RT session with probability P_{RT} or an NRT session with probability $P_{NRT} = 1 - P_{RT}$. The mean durations of the RT and NRT sessions are $1/\mu_{RT}$ and $1/\mu_{NRT}$ respectively.

6.4 Mobility Model

In a 3G-WLAN interworking network, a cellular cell is usually called a *3G-WLAN interworking cell*, and is generally overlaid with one or more WLAN cells. Therefore, the mobility model for a mobile node should characterise its residence times not only in the whole 3G-WLAN interworking cell but also in the different RAT areas within the 3G-WLAN interworking cell. To this end, the mobility of the mobile node in the 3G-WLAN interworking network is modelled as a CTMC which has the *Coxian* structure.

A K-phase Coxian structure is composed of a series of K exponentially distributed phases (or states) and an absorbing phase. Phase i either enters the phase $i+1$ or enters the absorbing phase with pre-defined probabilities. Phase K enters the absorbing phase with the probability of 1. A Coxian distribution is then defined as the distribution from the first phase until absorption, and it has an important property that it can arbitrarily closely approximate any probability distribution [122]. By letting the phases represent the position of the mobile node in a 3G-WLAN interworking cell in terms of the RAT area, the transitions of the Coxian structure can capture various tracks that the mobile node may traverse in a 3G-WLAN interworking cell [123].

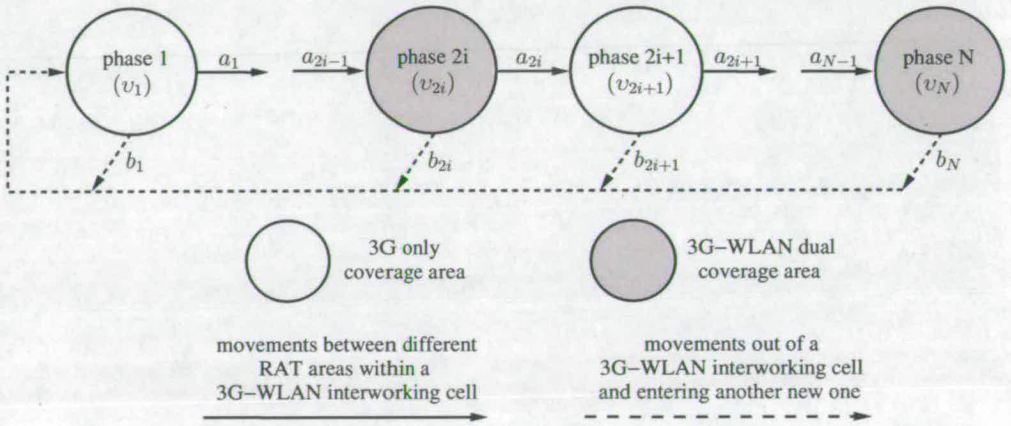
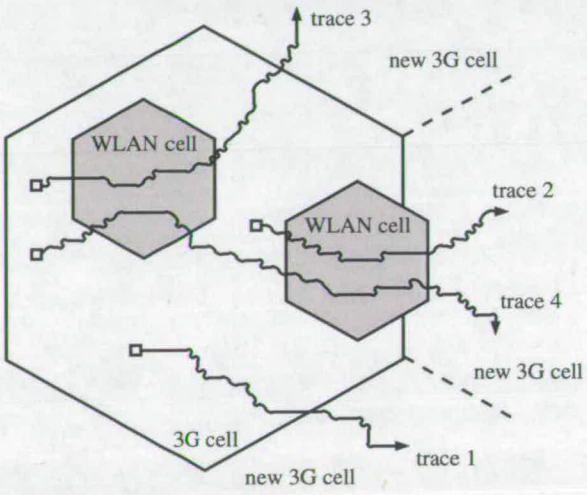


Figure 6.2: *A mobility model with the Coxian structure*

Since the ordinary Coxian structure contains an absorbing state whereas the focus of this work is on the steady state performance of a mobile node, the absorbing state is omitted in the mobility model. The mobility model with the Coxian structure is assumed to have an even number (N) of phases, and it is shown in Figure 6.2. The odd phases $2i-1$ and the even phases $2i$ represent the mobile node being in the *3G only coverage areas* and in the *3G-WLAN dual coverage areas* respectively, where $i = 1, 2, \ldots, N/2$. Without ambiguity, the above two types of areas are called a *3G area* and a *3G-WLAN area* respectively. Two important assumptions are made in this work: Firstly, WLAN cells are assumed not to overlap with each other. Secondly, a WLAN cell which overlaps with two adjacent cellular cells is considered to belong to *both* cellular cells. With the second assumption, the starting point of the track of the mobile node in a 3G-WLAN interworking cell is always the 3G area.

The state transition diagram shown in Figure 6.2 captures the movement of

a mobile node within and across the 3G-WLAN interworking cells: Movements across different RAT areas *within* a 3G-WLAN interworking cell are captured by the transitions between neighbouring phases. Movements *out of* the 3G-WLAN interworking cell are captured by the transitions from interim phases back to the first phase, and mean the mobile node enters the 3G area of *another new* 3G-WLAN interworking cell. For example, all the four tracks of the mobile node shown in Figure 6.3 can be captured by the CTMC mobility model.



trace 1: phase 1 > phase 1
trace 2: phase 1 > phase 2 > phase 1
trace 3: phase 1 > phase 2 > phase 3 > phase 1
trace 4: phase 1 > phase 2 > phase 3 > phase 4 > phase 1

Figure 6.3: *Examples of a mobile node's tracks in a 3G-WLAN interworking cell*

Owing to the Coxian structure of the mobility model, the sojourn time of the mobile node in the 3G-WLAN interworking cell, i.e., the time spent by the mobile node traversing a series of phases until going back to the first phase, still follows the Coxian distribution. To make a more accurate mobility model, each phase can be further modelled as a CTMC with the Coxian structure to approximate the sojourn time distribution in that RAT area. To simplify the mobility model, each phase of the mobility model, e.g. phase k, is assumed to be exponentially distributed with rate v_k, and the probabilities of branching to the next and the first phases are a_k and b_k respectively. The number of phases, the rate and the branching probabilities of each phase can be estimated from field data using algorithms presented in [122, 124].

6.5 PEPA Models of Network Selection Strategies

In this section, PEPA models that describe the behaviour of a mobile node using different NSSs are presented. Each strategy has its corresponding PEPA model, and each model consists of three PEPA components that capture the session, mobility and network selection behaviour of the mobile node. The same traffic and mobility patterns of the mobile node are used in each model in order that the results depend only on the characteristics of different selection strategies. In the following subsections, the PEPA components for the session and handover behaviour of the mobile node are presented first, followed by the PEPA components corresponding to each NSS.

6.5.1 PEPA Component for Traffic Model

A mobile node's session behaviour is modelled by the component *SESSION*. According to Section 6.4, it can be in an idle or an engaged state.

6.5.1.1 Idle State of Component *SESSION*

In state $SESSION_{Idle}$, the component *SESSION* can carry out two sets of new session request activities, depending on the position of the mobile node in the 3G-WLAN interworking cell.

- When the mobile node is in the 3G area, new session requests are generated at the rate of λ and all of them are submitted to the 3G RAN (3GRAN). The new session request activities are classified by the session type (activities $session\_request_{NRT}$ and $session\_request_{RT}$) and they are generated with the probabilities P_{NRT} and P_{RT} respectively.

- When the mobile node is in the 3G-WLAN area, the new session request activities are further classified by the RAN the requests are submitted to. For example, the activities $session\_request^{C}_{NRT}$ and $session\_request^{W}_{RT}$ mean that an NRT session request is submitted to the 3GRAN and an RT session request is submitted to the WLAN RAN (WRAN) respectively. The RAN to which the request is submitted is determined by the PEPA component for the NSS, which will be discussed in Section 6.5.3.

- Once the new session requests are admitted, the component *SESSION* goes to one of the engaged states. The probability of admitting new session requests is

reflected in the cooperation between the component *SESSION* and the PEPA component for mobility, which will be discussed in Section 6.5.2.[1] A new session request may be blocked by the 3GRAN and WRAN, and in this case the component *SESSION* remains in the idle state. Since the activities corresponding to blocking new session requests are self-transitions and have no effect on the equilibrium probability of the system states, they are omitted and not defined in the idle state.

The idle state of the component *SESSION* is defined as:

$$\begin{aligned} SESSION_{Idle} \overset{def}{=} & (session\_request_{NRT},\ P_{NRT} * \lambda).SESSION_{NRT} \\ & + (session\_request_{RT},\ P_{RT} * \lambda).SESSION_{RT} \\ & + (session\_request^{C}_{NRT},\ P_{NRT} * \lambda).SESSION_{NRT} \\ & + (session\_request^{W}_{NRT},\ P_{NRT} * \lambda).SESSION_{NRT} \\ & + (session\_request^{C}_{RT},\ P_{RT} * \lambda).SESSION_{RT} \\ & + (session\_request^{W}_{RT},\ P_{RT} * \lambda).SESSION_{RT} \end{aligned}$$

6.5.1.2 Engaged States of Component *SESSION*

Depending on the type of a session, the component *SESSION* can be either in the state $SESSION_{NRT}$ or in the state $SESSION_{RT}$. When the session is completed (activities $session_{NRT}$ and $session_{RT}$), or is dropped during a handover (activities *HHO_block* and *VHO_block*), the component *SESSION* goes back to the idle state. The engaged states of the component *SESSION* are defined as:

$$\begin{aligned} SESSION_{NRT} \overset{def}{=} & (session_{NRT},\ \mu_{NRT}).SESSION_{Idle} \\ & + (VHO\_block,\ \top).SESSION_{Idle} \\ & + (HHO\_block,\ \top).SESSION_{Idle} \\ SESSION_{RT} \overset{def}{=} & (session_{RT},\ \mu_{RT}).SESSION_{Idle} \\ & + (VHO\_block,\ \top).SESSION_{Idle} \\ & + (HHO\_block,\ \top).SESSION_{Idle} \end{aligned}$$

6.5.2 PEPA Component for Mobility Model

A mobile node's handover behaviour is captured by the component *MN*. Moreover, the component *MN* also expresses the session and network selection behaviour

[1]Therefore, the activity rates of successful new session requests are not $P_{NRT} * \lambda$ and $P_{RT} * \lambda$ and they are determined by the cooperation.

of the mobile node, but they are *controlled* by the corresponding PEPA components. Like the component *SESSION*, the component *MN* can be in an idle or an engaged state, and with the component for the NSS on the network selection activities.

6.5.2.1 Idle States of Component *MN*

The idle states of the component *MN* imply that the mobile node is not communicating and they are characterised by the position of the mobile node in the 3G-WLAN interworking cell. The subscripts $2i-1$ and $2i$ ($i = 1, 2, \ldots, N/2$) are used to denote which phase of the mobility model the mobile node is currently in. The idle states of the component *MN* cooperate with the component *SESSION* on different sets of new session request activities according to the mobile node's position.

- In state MN^{Idle}_{2i-1}, the mobile node is in the 3G area. It submits new session requests to the 3GRAN (activities $session\_request_{NRT}$ and $session\_request_{RT}$). The probability of admitting a new session request in the 3GRAN is reflected in the parameter P^C_{NA}. For example, by the cooperation between MN^{Idle}_{2i-1} and $SESSION_{Idle}$, the rate of the activity $session\_request_{NRT}$ is actually $P^C_{NA} * P_{NRT} * \lambda$, rather than $P_{NRT} * \lambda$.

- In state MN^{Idle}_{2i}, the mobile node is in the 3G-WLAN area. It submits new session requests to different RANs according to its selection strategy (e.g. activities $session\_request^C_{NRT}$ and $session\_request^W_{RT}$). The probability of admitting a new session request to the WRAN is P^W_{NA}. The probabilities of selecting 3GRAN and WRAN are P_C and P_W respectively and they are defined in the PEPA component for NSS. For example, the rate of the activity $session\_request^W_{RT}$ is actually $P^W_{NA} * P_W * P_{RT} * \lambda$.

- Once new session requests are admitted, the component *MN* goes into an engaged state. As for the component *SESSION*, the activities corresponding to blocking new session requests are omitted since they are self-transitions.

- The mobile node can stay in the idle state and just move within and across the 3G-WLAN interworking cells. The activity $move_{m,n}$ represents the movement of the mobile node from phase m to phase n of its mobility model.

The idle states of the component MN are defined as ($i = 1, 2, \ldots, N/2$):

$$
\begin{aligned}
MN^{Idle}_{2i-1} \stackrel{def}{=} & \ (session\_request_{NRT},\ P^{C}_{NA} * \top).MN^{C}_{2i-1} \\
& + (session\_request_{RT},\ P^{C}_{NA} * \top).MN^{C}_{2i-1} \\
& + (move_{2i-1,2i},\ a_{2i-1} * \upsilon_{2i-1}).MN^{Idle}_{2i} \\
& + (move_{2i-1,1},\ b_{2i-1} * \upsilon_{2i-1}).MN^{Idle}_{1} \\
MN^{Idle}_{2i} \stackrel{def}{=} & \ (session\_request^{C}_{NRT},\ P^{C}_{NA} * \top).MN^{C}_{2i} \\
& + (session\_request^{C}_{RT},\ P^{C}_{NA} * \top).MN^{C}_{2i} \\
& + (session\_request^{W}_{NRT},\ P^{W}_{NA} * \top).MN^{W}_{2i} \\
& + (session\_request^{W}_{RT},\ P^{W}_{NA} * \top).MN^{W}_{2i} \\
& + (move_{2i,2i+1},\ a_{2i} * \upsilon_{2i}).MN^{Idle}_{2i+1} \\
& + (move_{2i,1},\ b_{2i} * \upsilon_{2i}).MN^{Idle}_{1} \\
MN^{Idle}_{N} \stackrel{def}{=} & \ (session\_request^{C}_{NRT},\ P^{C}_{NA} * \top).MN^{C}_{N} \\
& + (session\_request^{C}_{RT},\ P^{C}_{NA} * \top).MN^{C}_{N} \\
& + (session\_request^{W}_{NRT},\ P^{W}_{NA} * \top).MN^{W}_{N} \\
& + (session\_request^{W}_{RT},\ P^{W}_{NA} * \top).MN^{W}_{N} \\
& + (move_{N,1},\ b_{N} * \upsilon_{N}).MN^{Idle}_{1}
\end{aligned}
$$

6.5.2.2 Engaged States of Component *MN*

The engaged states of the component MN imply that the mobile node is communicating and they are characterised by both the position of the mobile node in the 3G-WLAN interworking cell and the RAN it is currently connected to. The superscripts C and W are used to denote the mobile node is using the 3GRAN and WRAN respectively. The engaged states of the component MN cooperate with the component $SESSION$ on the session holding activities and with the component for the NSS on the network selection activities.

- In state MN^{C}_{2i-1}, the mobile node is in the 3G area and is connected to the 3GRAN. When it moves into another 3G-WLAN interworking cell (i.e. from phase $2i-1$ to phase 1), it performs a horizontal handover (HHO) (activity $HHO_{2i-1,1}$). When it moves into the 3G-WLAN area of the same cell (i.e. from phase $2i-1$ to phase $2i$), it either performs a VHO to the WRAN (activity $VHO_{2i-1,2i}$), or performs no handover (NHO) and keeps its connection to the 3GRAN (activity $NHO_{2i-1,2i}$). Which type of handover is performed is decided

by its NSS and is controlled by the PEPA component for NSS.[II]

- In state MN^C_{2i}, the mobile node is in the 3G-WLAN area and is connected to the 3GRAN. When it moves into another 3G-WLAN interworking cell (i.e. from phase $2i$ to phase 1), it performs a HHO (activity $HHO_{2i,1}$). When it moves into the 3G area of the same 3G-WLAN interworking cell (i.e. from phase $2i$ to phase $2i+1$), no handover is required (activity $NHO_{2i,2i+1}$).

- In state MN^W_{2i}, the mobile node is in the 3G-WLAN area and is connected to the WRAN. It always performs a VHO (activities $VHO_{2i,1}$ and $VHO_{2i,2i+1}$) when it moves out of the 3G-WLAN area (i.e. from phase $2i-1$ to phase 1 and from phase $2i-1$ to phase $2i$).

  To help understand the above different types of handovers, Figure 6.4 illustrates the handover related transitions between the engaged states.

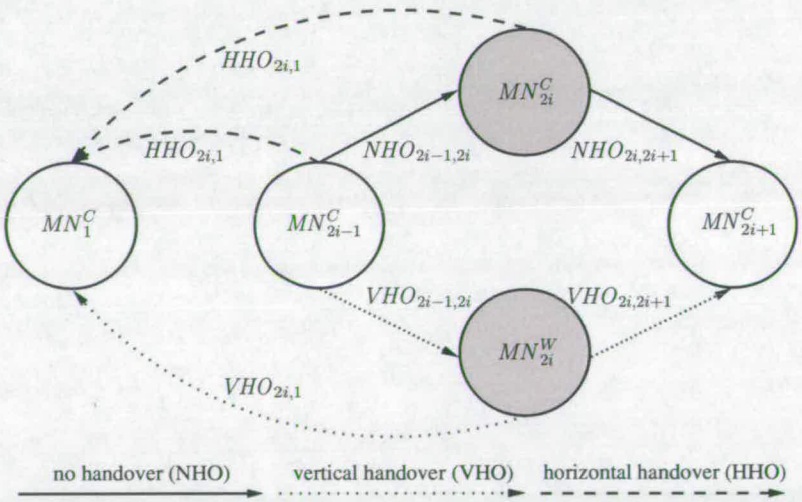


Figure 6.4: *Different types of handovers between the engaged states*

- When the mobile node finishes its session (activities $session_{NRT}$ and $session_{RT}$), the component MN returns to an idle state. Generally, the NRT sessions are aware of the different data rates provided by different RATs. Therefore, the factors R^C_{NRT} and R^W_{NRT} are used to adjust the duration of the NRT sessions when the mobile node is connected to the 3GRAN and WRAN respectively. For example, the duration of the NRT session in the 3GRAN is $1/(R^C_{NRT} * \mu_{NRT})$.

[II]Note that since the network selection only happens when the mobile node moves from the 3G area into the 3G-WLAN area, only $NHO_{2i-1,2i}$, $VHO_{2i-1,2i}$ and *VHO_block* are *network selection related* activities during a handover.

- When the horizontal and vertical handover requests of the mobile node are blocked (activities *HHO_block* and *VHO_block*), the component *MN* also goes back to an idle state. The probabilities of admitting and blocking handover requests in the 3GRAN and WRAN are P^C_{HA}, P^C_{HB}, P^W_{HA}, and P^W_{HB} respectively.

The engaged states of the component *MN* are defined as ($i = 1, 2, \ldots, N/2$):

$$
\begin{aligned}
MN^C_{2i-1} \stackrel{def}{=}\ & (session_{NRT},\ R^C_{NRT} * \top).MN^{Idle}_{2i-1} \\
& + (session_{RT},\ \top).MN^{Idle}_{2i-1} \\
& + (HHO_{2i-1,1},\ P^C_{HA} * b_{2i-1} * \upsilon_{2i-1}).MN^C_1 \\
& + (HHO\_block,\ P^C_{HB} * b_{2i-1} * \upsilon_{2i-1}).MN^{Idle}_1 \\
& + (NHO_{2i-1,2i},\ a_{2i-1} * \upsilon_{2i-1}).MN^C_{2i} \\
& + (VHO_{2i-1,2i},\ P^W_{HA} * a_{2i-1} * \upsilon_{2i-1}).MN^W_{2i} \\
& + (VHO\_block,\ P^W_{HB} * a_{2i-1} * \upsilon_{2i-1}).MN^{Idle}_{2i} \\
MN^C_{2i} \stackrel{def}{=}\ & (session_{NRT},\ R^C_{NRT} * \top).MN^{Idle}_{2i} \\
& + (session_{RT},\ \top).MN^{Idle}_{2i} \\
& + (HHO_{2i,1},\ P^C_{HA} * b_{2i} * \upsilon_{2i}).MN^C_1 \\
& + (HHO\_block,\ P^C_{HB} * b_{2i} * \upsilon_{2i}).MN^{Idle}_1 \\
& + (NHO_{2i,2i+1},\ a_{2i} * \upsilon_{2i}).MN^C_{2i+1} \\
MN^W_{2i} \stackrel{def}{=}\ & (session_{NRT},\ R^W_{NRT} * \top).MN^{Idle}_{2i} \\
& + (session_{RT},\ \top).MN^{Idle}_{2i} \\
& + (VHO_{2i,1},\ P^C_{HA} * b_{2i} * \upsilon_{2i}).MN^C_1 \\
& + (VHO\_block,\ P^C_{HB} * b_{2i} * \upsilon_{2i}).MN^{Idle}_1 \\
& + (VHO_{2i,2i+1},\ P^C_{HA} * a_{2i} * \upsilon_{2i}).MN^C_{2i+1} \\
& + (VHO\_block,\ P^C_{HB} * a_{2i} * \upsilon_{2i}).MN^{Idle}_{2i+1} \\
MN^C_N \stackrel{def}{=}\ & (session_{NRT},\ R^C_{NRT} * \top).MN^{Idle}_N \\
& + (session_{RT},\ \top).MN^{Idle}_N \\
& + (HHO_{N,1},\ P^C_{HA} * \upsilon_N).MN^C_1 \\
& + (HHO\_block,\ P^C_{HB} * \upsilon_N).MN^{Idle}_1 \\
MN^W_N \stackrel{def}{=}\ & (session_{NRT},\ R^W_{NRT} * \top).MN^{Idle}_N \\
& + (session_{RT},\ \top).MN^{Idle}_N \\
& + (VHO_{N,1},\ P^C_{HA} * \upsilon_N).MN^C_1 \\
& + (VHO\_block,\ P^C_{HB} * \upsilon_N).MN^{Idle}_1
\end{aligned}
$$

6.5.3 PEPA Component and System Definition for Each Network Selection Strategy

A mobile node's network selection behaviour is controlled by the component NS, and a different component NS is required for each strategy. The component NS is designed to synchronise with the component $SESSION$ and the component MN on the network selection related activities. In this way, the mobile node's choice of a certain RAN can easily be controlled by enabling and disabling the synchronisation of the corresponding activities.

Different NSSs are classified into two groups: non-deterministic and deterministic. Non-deterministic strategies choose the RAN according to some on-line measures as discussed in Chapter 2. The resultant RAN is not necessarily the same at every decision. On the other hand, deterministic strategies choose the RAN based on the traffic type, the user preference, etc. In this work, PEPA models of three types of selection strategies are built, namely general, WLAN-first and service-based. The general NSS model is for the common non-deterministic strategies and it embeds randomness in network selection. The WLAN-first and service-based NSSs models are for the specific deterministic strategies as their names suggest. The component NS and the system definition for each type of strategy are described in the following subsections.

6.5.3.1 General Network Selection Strategy

Using the general strategy, the mobile node chooses the 3GRAN and WRAN with non-zero probabilities P_C and P_W respectively. One example is the random strategy which chooses the 3GRAN and the WRAN with equal probabilities. Another example is the strategy that is based on the relative received signal strength (RRSS) [116]. No matter how the non-deterministic strategies are designed, the PEPA component for the general network selection NS^G should enable selecting both RANs for new and

handover sessions. The component NS^G is defined as:

$$\begin{aligned} NS^G \stackrel{def}{=} & (session\_request^C_{NRT},\ P_C * \top).NS^G \\ & + (session\_request^C_{RT},\ P_C * \top).NS^G \\ & + (session\_request^W_{NRT},\ P_W * \top).NS^G \\ & + (session\_request^W_{RT},\ P_W * \top).NS^G \\ & + (NHO_{2i-1,2i},\ P_C * \top).NS^G \\ & + (VHO_{2i-1,2i},\ P_W * \top).NS^G \\ & + (VHO\_block,\ P_W * \top).NS^G \end{aligned}$$

System Definition: The network selection for a new session request is implemented by synchronising the components NS^G, *SESSION* and *MN* on the activities $session\_request^C_{NRT}$, $session\_request^C_{RT}$, $session\_request^W_{NRT}$, and $session\_request^W_{RT}$. The network selection for a handover session request is implemented by cooperation between the components NS^G and *MN* on the activities $NHO_{2i-1,2i}$, $VHO_{2i-1,2i}$, and *VHO_block*. The other session related activities of the component *MN* are controlled by its synchronisation with the component *SESSION*. Therefore, the PEPA model of the general NSS is constructed as:

$$NSS^G \stackrel{def}{=} SESSION_{Idle} \underset{L_1}{\bowtie} MN_1^{Idle} \underset{L_2}{\bowtie} NS^G,$$

where

$$\begin{aligned} L_1 = \Big\{ & session\_request^C_{NRT}, session\_request^C_{RT}, session\_request^W_{NRT}, \\ & session\_request^W_{RT}, session\_request_{NRT}, session\_request_{RT}, \\ & session_{NRT}, session_{RT}, HHO\_block, VHO\_block \Big\}, \\ L_2 = \Big\{ & session\_request^C_{NRT}, session\_request^C_{RT}, session\_request^W_{NRT}, \\ & session\_request^W_{RT}, NHO_{2i-1,2i}, VHO_{2i-1,2i}, VHO\_block \Big\}. \end{aligned}$$

6.5.3.2 WLAN-first Network Selection Strategy

Using the WLAN-first strategy, the mobile node always chooses the WRAN whenever it is available. WRAN is usually preferred because of its high bandwidth, small delay and low cost. This strategy can be implemented by disabling the activities corresponding to selecting the 3GRAN for both new and handover sessions. The PEPA

component for WLAN-first network selection NS^{WF} is defined as:

$$\begin{aligned} NS^{WF} \stackrel{def}{=} & (session\_request^{W}_{NRT},\ P_W * \top).NS^{WF} \\ & + (session\_request^{W}_{RT},\ P_W * \top).NS^{WF} \\ & + (VHO_{2i-1,2i},\ P_W * \top).NS^{WF} \\ & + (VHO\_block,\ P_W * \top).NS^{WF} \end{aligned}$$

System Definition: The system definition of the WLAN-first NSS use the same structure and cooperation sets as those of the general NSS. Since the component NS^{WF} does not allow the choice of the 3GRAN, WLAN is always chosen. The PEPA model of the WLAN-first NSS is constructed as:

$$NSS^{WF} \stackrel{def}{=} SESSION_{Idle} \underset{L_1}{\bowtie} MN_1^{Idle} \underset{L_2}{\bowtie} NS^{WF},$$

where

$$\begin{aligned} L_1 = \Big\{ & session\_request^{C}_{NRT}, session\_request^{C}_{RT}, session\_request^{W}_{NRT}, \\ & session\_request^{W}_{RT}, session\_request_{NRT}, session\_request_{RT}, \\ & session_{NRT}, session_{RT}, HHO\_block,\ VHO\_block \Big\}, \\ L_2 = \Big\{ & session\_request^{C}_{NRT}, session\_request^{C}_{RT}, session\_request^{W}_{NRT}, \\ & session\_request^{W}_{RT}, NHO_{2i-1,2i},\ VHO_{2i-1,2i},\ VHO\_block \Big\}. \end{aligned}$$

6.5.3.3 Service-based Network Selection Strategy

Using the service-based strategy, the mobile node chooses the RAN according to the type of its ongoing session. For example, the mobile node may choose the WRAN for NRT sessions and 3GRAN for RT sessions. This choice is based on the fact that the NRT sessions can take advantage of the higher data rate provided by WRAN and the RT sessions will experience less handovers when choosing 3GRAN.

Since the mobile node makes its decision based on the session type, it would be better to let the component *SESSION* control the network selection behaviour. For that reason, the engaged states of the component *SESSION* are modified to implement service-based network selection during *handover*. For the NRT sessions, since the mobile node always performs a VHO from the 3GRAN to the WRAN, the activity $VHO_{2i-1,2i}$ is enabled. On the other hand, since the RT sessions always use the 3GRAN

and no vertical handover is required, the activity $NHO_{2i-1,2i}$ is enabled and the activity *VHO_block* is not needed. The modified component *SESSION* for the service-based strategy, which is renamed as $SESSION^{SB}$, is defined as:

$$\begin{aligned}
SESSION^{SB}_{Idle} &\stackrel{def}{=} (session_request_{NRT},\ P_{NRT} * \lambda).SESSION^{SB}_{NRT} \\
&\quad + (session_request_{RT},\ P_{RT} * \lambda).SESSION^{SB}_{RT} \\
&\quad + (session_request^{C}_{NRT},\ P_{NRT} * \lambda).SESSION^{SB}_{NRT} \\
&\quad + (session_request^{W}_{NRT},\ P_{NRT} * \lambda).SESSION^{SB}_{NRT} \\
&\quad + (session_request^{C}_{RT},\ P_{RT} * \lambda).SESSION^{SB}_{RT} \\
&\quad + (session_request^{W}_{RT},\ P_{RT} * \lambda).SESSION^{SB}_{RT} \\
SESSION^{SB}_{NRT} &\stackrel{def}{=} (session_{NRT},\ \mu_{NRT}).SESSION^{SB}_{Idle} \\
&\quad + (VHO_{2i-1,2i},\ P_W * \top).SESSION^{SB}_{NRT} \\
&\quad + (VHO_block,\ P_W * \top).SESSION^{SB}_{Idle} \\
&\quad + (HHO_block,\ \top).SESSION^{SB}_{Idle} \\
SESSION^{SB}_{RT} &\stackrel{def}{=} (session_{RT},\ \mu_{RT}).SESSION^{SB}_{Idle} \\
&\quad + (NHO_{2i-1,2i},\ P_C * \top).SESSION^{SB}_{RT} \\
&\quad + (HHO_block,\ \top).SESSION^{SB}_{Idle}
\end{aligned}$$

The network selection for a new session request is still implemented in the component NS^{SB} by enabling corresponding activities.[III] The component NS^{SB} is defined as:

$$\begin{aligned}
NS^{SB} &\stackrel{def}{=} (session_request^{C}_{RT},\ P_C * \top).NS^{SB} \\
&\quad + (session_request^{W}_{NRT},\ P_W * \top).NS^{SB}
\end{aligned}$$

System Definition: The system definition of the service-based strategy has the same structure as the previous two strategies but its cooperation sets are different. The PEPA model of the service-based NSS is constructed as:

$$NSS^{SB} \stackrel{def}{=} SESSION^{SB}_{Idle} \underset{L_3}{\bowtie} MN^{Idle}_{1} \underset{L_4}{\bowtie} NS^{SB},$$

[III] In fact, the network selection for a new session request can also be implemented by modifying the idle state of the component *SESSION*. The component NS^{SB} is deliberately used so that the model has the same structure as the other models.

where

$$L_3 = \Big\{ session\_request^{C}_{NRT}, session\_request^{C}_{RT}, session\_request^{W}_{NRT}, session\_request^{W}_{RT}, session\_request_{NRT}, session\_request_{RT}, session_{NRT}, session_{RT}, NHO_{2i-1,2i}, VHO_{2i-1,2i}, HHO\_block, VHO\_block \Big\},$$

$$L_4 = \Big\{ session\_request^{C}_{NRT}, session\_request^{C}_{RT}, session\_request^{W}_{NRT}, session\_request^{W}_{RT} \Big\}.$$

6.5.4 System States of the PEPA Models

The PEPA models of each NSS are structured in the same way and the system states of each model are captured by the component *MN* and the component *SESSION*. For example, the PEPA model of the general NSS has the system state:

$$\left(SESSION_{NRT} | MN^{W}_{2i} | NS^{G}\right).$$

In this chapter a system state of a PEPA model is denoted as $s^{A,B}_{k}$, where k, A and B represent the mobile node's phase of its mobility model, the RAN it is connected to, and the type of the session it is engaged in respectively ($k = 1, 2, \cdots, N$). For example, $s^{C,RT}_{3}$ means the mobile node is in phase 3 and is connected to the 3GRAN for an RT session. Moreover, s^{C}_{k} and s^{W}_{k} are the unions of system states and they are defined as $s^{C}_{k} = s^{C,NRT}_{k} \cup s^{C,RT}_{k}$ and $s^{W}_{k} = s^{W,NRT}_{k} \cup s^{W,RT}_{k}$. Note that the system states of all the models are the *feasible* combinations of the state of each component of that model. As in previous chapters, the feasibility of system states of all the models are checked. For example, the model of the WLAN-first NSS does not have the states $s^{C,RT}_{2l}$ and $s^{C,NRT}_{2l}$, and the model of the service-based NSS does not have the states $s^{W,RT}_{2l}$ and $s^{C,NRT}_{2l}$, where $l = 1, 2, \cdots, N/2$.

6.5.5 Performance Measures

Three performance measures, namely average throughput, RAN blocking probability and handover rate are investigated in this work.

- **Average Throughput**: The average throughput is defined as the mean data rate that can be achieved by a mobile node during its communication. To derive this measure, the first step is to obtain the percentages of time the mobile node

spends using different RATs for different types of sessions. Therefore, four types of engaged times can be defined as follows:

$$T_{C,NRT} = \sum_{i=1}^{N} \pi(s_i^{C,NRT}), \tag{6.1}$$

$$T_{C,RT} = \sum_{i=1}^{N} \pi(s_i^{C,RT}), \tag{6.2}$$

$$T_{W,NRT} = \sum_{i=1}^{N/2} \pi(s_{2i}^{W,NRT}), \tag{6.3}$$

$$T_{W,RT} = \sum_{i=1}^{N/2} \pi(s_{2i}^{W,RT}), \tag{6.4}$$

where $\pi(s_k^{A,B})$ is the equilibrium probability of system state $s_k^{A,B}$. Then the total percentage of time the mobile node is in the engaged states is:

$$T_{Engaged} = T_{C,NRT} + T_{C,RT} + T_{W,NRT} + T_{W,RT}. \tag{6.5}$$

Based on the above definitions, the average throughput is calculated as the weighted sum of the proportions of the different engaged times to the total engaged time, where the weights are the corresponding data rates of different RATs. That is:

$$\begin{aligned} THP = & D_{NRT}^{C} * \frac{T_{C,NRT}}{T_{Engaged}} + D_{RT}^{C} * \frac{T_{C,RT}}{T_{Engaged}} \\ & + D_{NRT}^{W} * \frac{T_{W,NRT}}{T_{Engaged}} + D_{RT}^{W} * \frac{T_{W,RT}}{T_{Engaged}}, \end{aligned} \tag{6.6}$$

where D_{NRT}^{C}, D_{RT}^{C}, D_{NRT}^{W} and D_{RT}^{W} are the data rates that can be achieved by the mobile node when it uses the 3G RAT (3GRAT) and WLAN RAT (WRAT) for NRT and RT sessions respectively.

- **RAN Blocking Probability**: Although the blocking probabilities of different RANs can be regarded as independent input parameters to the PEPA models, in this work an approach which utilises the PEPA models to derive the blocking probabilities is presented.

- **Handover Rate**: The handover rate is defined as the mean number of handover

attempts performed by the mobile node per unit time.

6.6 Derivation of RAN Blocking Probability and Handover Rate

As discussed in Section 6.5.5, the blocking probabilities of the 3GRAN and WRAN are not considered as input parameters. Instead, for a certain type of NSS, they are derived from the interaction between the PEPA model and a resource consumption model corresponding to that type of NSS. In this procedure, the horizontal and vertical handover rates of the mobile node are obtained at the same time. In the following subsections, the mathematical expressions of the RAN blocking probability and handover rate are presented first, followed by an iterative method to derive them.

6.6.1 RAN Blocking Probability

To derive the RAN blocking probability, a two-dimensional continuous-time Markov chain (2D-CTMC) is used to model the resource consumption of a 3G-WLAN interworking cell. The state of the 2D-CTMC is denoted by two nonnegative integers (c, w), where c and w are the numbers of engaged users in the 3GRAN and WRAN respectively. For WLAN cells which overlap with two adjacent cellular cells, their resources are assumed to be shared by both 3G-WLAN interworking cells. That is, the changes in the number of users of these WLAN cells are reflected in the changes of the states of their spanning 3G-WLAN interworking cells. As shown in Figure 6.5, there are five types of events that change the state of the 2D-CTMC and they are described as follows:

- **type 1**: New sessions requests are generated in the 3GRAN and the WRAN, and their rates are denoted as λ_C^n and λ_W^n respectively.
- **type 2**: Sessions are completed and resources in the 3GRAN and the WRAN are released, and their rates are denoted as μ_C and μ_W respectively.
- **type 3**: Sessions are *internally* handed over *between* the 3GRAN and the WRAN. Their rates are denoted as r_{C-W}^{intra} and r_{W-C}^{intra} respectively.
- **type 4**: Sessions are *externally* handed over *out of* the 3GRAN and the WRAN. Their rates are denoted as r_{C-C}^{inter} and r_{W-C}^{inter} respectively.

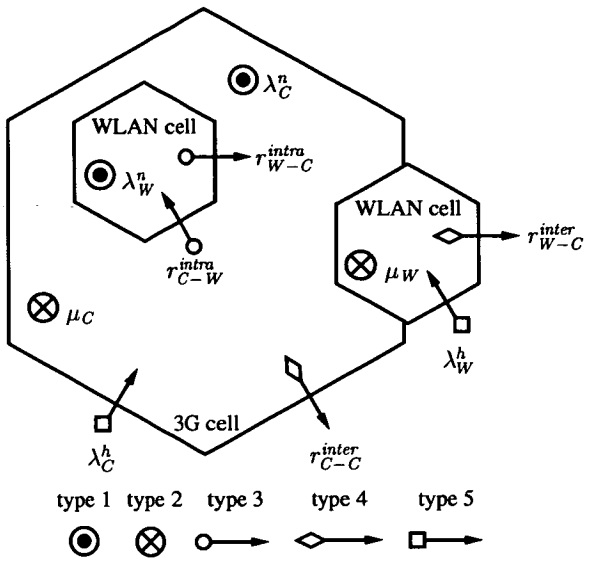


Figure 6.5: *Five types of events that change the state of the 2D-CTMC*

- **type 5**: Sessions are *externally* handed over *into* the 3GRAN and WRAN and their rates are denoted as λ_C^h and λ_W^h respectively.

According to the events described above, the state transition diagram of the 2D-CTMC can be generated. Figure 6.6 shows the outward transitions of a non-boundary state (c, w) of the 2D-CTMC. The whole state transition diagram of the 2D-CTMC can be constructed straightforwardly.

For each type of NSS, the rates of the transitions corresponding to the five types of events are calculated as follows:

- **type 1**: Assume that the mobile nodes in the 3G-WLAN interworking cell are uniformly distributed and let A_C and A_W denote the coverage percentage of 3G area and the 3G-WLAN area respectively. For the general and WLAN-first strategies, λ_C^n and λ_W^n are calculated as:

$$\lambda_C^n = A_C * \Lambda^n + P_C * A_W * \Lambda^n, \tag{6.7}$$

$$\lambda_W^n = P_W * A_W * \Lambda^n, \tag{6.8}$$

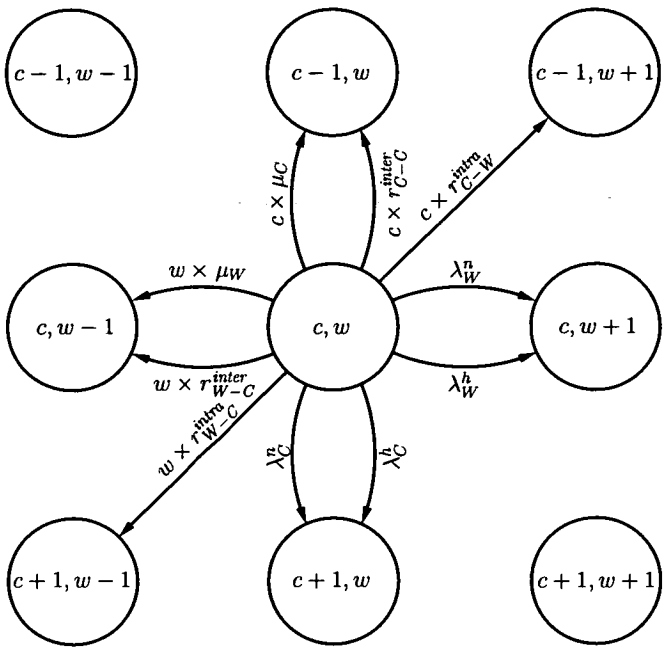


Figure 6.6: *Outward transitions of a non-marginal state of the 2D-CTMC*

and for the service-based strategy they are calculated as:

$$\lambda_C^n = A_C * \Lambda^n + P_{RT} * A_W * \Lambda^n, \tag{6.9}$$

$$\lambda_W^n = P_{NRT} * A_W * \Lambda^n, \tag{6.10}$$

where Λ^n is the arrival rate of the new session requests of the whole 3G-WLAN interwork cell.

- **type 2**: For the general and WLAN-first strategies, the resources of the 3GRAN and the WRAN can be used by both NRT and RT users. Therefore in both RANs the probabilities that a session is NRT or RT are P_{NRT} or P_{RT} respectively, and the resources holding time in the 3GRAN ($1/\mu_C$) and the WRAN ($1/\mu_W$) are calculated as:

$$\frac{1}{\mu_C} = \frac{P_{NRT}}{R_{NRT}^C * \mu_{NRT}} + \frac{P_{RT}}{\mu_{RT}}, \tag{6.11}$$

$$\frac{1}{\mu_W} = \frac{P_{NRT}}{R_{NRT}^W * \mu_{NRT}} + \frac{P_{RT}}{\mu_{RT}}. \tag{6.12}$$

For the service-based strategy, resource consumption is a bit more complex. The resources of the WRAN can only be used by NRT users in the 3G-WLAN area,

whereas the resources of the 3GRAN can be used by RT users in the 3G-WLAN area and all the users in the 3G area. To simplify the analysis, it is assumed that the probabilities that a session in the 3GRAN is NRT and RT are still P_{NRT} and P_{RT} respectively. Accordingly, $1/\mu_C$ and $1/\mu_W$ are calculated as:

$$\frac{1}{\mu_C} \approx \frac{P_{NRT}}{R^C_{NRT} * \mu_{NRT}} + \frac{P_{RT}}{\mu_{RT}}, \tag{6.13}$$

$$\frac{1}{\mu_W} = \frac{1}{R^W_{NRT} * \mu_{NRT}}. \tag{6.14}$$

- **type 3 and 4**: r^{intra}_{C-W}, r^{intra}_{W-C}, r^{inter}_{C-C} and r^{inter}_{W-C} are derived by the approach which will be discussed in Section 6.6.3.
- **type 5**: λ^h_C and λ^h_W are regarded as input parameters.

Once the transition rates are obtained, the generator matrix of the 2D-CTMC can be generated and its equilibrium probability vector can be derived. Then the blocking probabilities of the 3GRAN and the WRAN are calculated as:

$$P^C_B = \sum_{\substack{c=N_C \\ 0 \leqslant w \leqslant N_W}} p(c,w), \tag{6.15}$$

$$P^W_B = \sum_{\substack{w=N_W \\ 0 \leqslant c \leqslant N_C}} p(c,w), \tag{6.16}$$

where $p(c,w)$ is the equilibrium probability of the state (c,w) of the 2D-CTMC, and N_C and N_W are the maximum number of 3G and WLAN users that can be supported in a 3G-WLAN interworking cell respectively.[IV] Note that a different 2D-CTMC is required for each NSS since their generator matrices are different. Eq. (6.15) and Eq. (6.16) are the general expressions for all types of selection strategies.

6.6.2 Handover Rate

Since the handover rate is defined as the mean number of handover attempts performed by the mobile node per unit time, it is actually the throughput of the handover activity. Therefore, the handover rate can be calculated using MRS and Eq. (3.3), where the rewards are equal to the activity rate of that type of handover and

[IV] The capacity of the 3G-WLAN interworking cell is a complex topic and depends on the user data rate requirements, signal-to-noise ratio, transmission power, etc [125]. The static capacity in terms of the number of users is used in order to simply the analysis.

they are associated with the system states in which that type of handover is enabled. For example, the activity rate of the internal vertical handover from WRAN to 3GRAN is $a_{2i} * \upsilon_{2i}$, and the mobile node performs this type of handover in the system states s_{2i}^{W}.

Four types of handovers are expressed by the mobility model and their corresponding handover rates are calculated as follows:

$$r_{C-C}^{inter} = \sum_{i=1}^{N/2} \left(b_{2i-1} * \upsilon_{2i-1} * \pi(s_{2i-1}^{C}) + b_{2i} * \upsilon_{2i} * \pi(s_{2i}^{C}) \right), \tag{6.17}$$

$$r_{W-C}^{inter} = \sum_{i=1}^{N/2} b_{2i} * \upsilon_{2i} * \pi(s_{2i}^{W}), \tag{6.18}$$

$$r_{C-W}^{intra} = \sum_{i=1}^{N/2} P_W * a_{2i-1} * \upsilon_{2i-1} * \pi(s_{2i-1}^{C}), \tag{6.19}$$

$$r_{W-C}^{intra} = \sum_{i=1}^{N/2} a_{2i} * \upsilon_{2i} * \pi(s_{2i}^{W}), \tag{6.20}$$

where $\pi(s_{2i-1}^{C})$, $\pi(s_{2i}^{C})$ and $\pi(s_{2i}^{W})$ are the equilibrium probabilities of the corresponding system states of the PEPA models.

6.6.3 An Iterative Method to Derive RAN Blocking Probability and Handover Rate

As discussed above, the RAN blocking probabilities $\boldsymbol{P_B} = [P_B^C, P_B^W]$ are calculated from the equilibrium probability vector of the 2D-CTMC model $\boldsymbol{p}(c, w)$, which requires the handover rates $\boldsymbol{R_H} = [r_{C-W}^{intra}, r_{W-C}^{intra}, r_{C-C}^{inter}, r_{W-C}^{inter}]$ be derived from the PEPA models. On the other hand, $\boldsymbol{R_H}$ are calculated from the equilibrium probability vector of the PEPA models $\pi(s_k^{A,B})$, which requires $\boldsymbol{P_B}$ be derived from the 2D-CTMC model. Therefore, for each type of NSS, its corresponding 2D-CTMC and PEPA models form a closed loop by exchanging $\boldsymbol{P_B}$ and $\boldsymbol{R_H}$, i.e.,

$$\begin{aligned} F_{CTMC}(\boldsymbol{R_H}): \quad \boldsymbol{R_H} &\xrightarrow{f_{CTMC}} \boldsymbol{p}(c,w) \xrightarrow{\text{(6.15) - (6.16)}} \boldsymbol{P_B}, \\ F_{PEPA}(\boldsymbol{P_B}): \quad \boldsymbol{P_B} &\xrightarrow{f_{PEPA}} \boldsymbol{\pi}(s_k^{A,B}) \xrightarrow{\text{(6.17) - (6.20)}} \boldsymbol{R_H}, \end{aligned}$$

where f_{CTMC} and f_{PEPA} denote the mathematical computation to obtain the equilibrium probability vectors of the 2D-CTMC and the PEPA models respectively. The two procedures to derive $\boldsymbol{P_B}$ and $\boldsymbol{R_H}$ are denoted as $F_{CTMC}(\boldsymbol{R_H})$ and $F_{PEPA}(\boldsymbol{P_B})$

respectively. To solve this implicit problem, an iterative method is designed and it is described in Algorithm 6.1.

Algorithm 6.1: An iterative method to derive $\boldsymbol{P}_B$ and $\boldsymbol{R}_H$

Input: all the required parameters
Output: $\boldsymbol{P}_B = [P_B^C, P_B^W]$, $\boldsymbol{R}_H = [r_{C-W}^{intra}, r_{W-C}^{intra}, r_{C-C}^{inter}, r_{W-C}^{inter}]$.

1 $conv = 0$; /* convergence tag */
2 $\epsilon = 10^{-30}$; /* convergence criteria */
3 $i = 1$; /* iteration counter */
4 $iter = 100$; /* maximum iteration number */
5 $\boldsymbol{P}_B^{(1)} = [0, 0]$; /* initialise $\boldsymbol{P}_B^{(1)}$ */
6 $\boldsymbol{R}_H^{(1)} = F_{PEPA}(\boldsymbol{P}_B^{(1)})$; /* calculate $\boldsymbol{R}_H^{(1)}$ using $\boldsymbol{P}_B^{(1)}$ */
7 **while** *(*$i \leq iter$*) or (*$conv = 0$*)* **do**
8 $\quad \boldsymbol{P}_B^{(i+1)} = F_{CTMC}(\boldsymbol{R}_H^{(i)})$; /* update $\boldsymbol{P}_B^{(i+1)}$ using $\boldsymbol{R}_H^{(i)}$ */
9 $\quad err = \max\{\boldsymbol{P}_B^{(i+1)} - \boldsymbol{P}_B^{(i)}\}$; /* calculate difference err */
10 $\quad$ **if** *(*$err > \epsilon$*)* **then**
11 $\quad\quad \boldsymbol{R}_H^{(i+1)} = F_{PEPA}(\boldsymbol{P}_B^{(i+1)})$; /* calculate $\boldsymbol{R}_H^{(i+1)}$ using $\boldsymbol{P}_B^{(i+1)}$ */
12 $\quad\quad \boldsymbol{R}_H = \boldsymbol{R}_H^{(i+1)}$; /* update $\boldsymbol{R}_H$ */
13 $\quad\quad \boldsymbol{P}_B = \boldsymbol{P}_B^{(i+1)}$; /* update $\boldsymbol{P}_B$ */
14 $\quad\quad i = i + 1$;
15 $\quad$ **else**
16 $\quad\quad conv = 1$; /* set the convergence tag */
17 $\quad\quad \boldsymbol{R}_H = \boldsymbol{R}_H^{(i)}$; /* update $\boldsymbol{R}_H$ */
18 $\quad\quad \boldsymbol{P}_B = \boldsymbol{P}_B^{(i)}$; /* update $\boldsymbol{P}_B$ */
19 $\quad$ **end**
20 **end**
21 **return** $\boldsymbol{P}_B, \boldsymbol{R}_H$;

Steps 1 to 4 are the necessary initialisations such as the convergence tag and convergence criteria etc. The initial values of $\boldsymbol{P}_B^{(1)}$ are set to 0 and the initial values of $\boldsymbol{R}_H^{(1)}$ are calculated using $\boldsymbol{P}_B^{(1)}$. Steps 7 to 20 carry out the iterative calculation. At the ith iteration, $\boldsymbol{R}_H^{(i)}$ is used to calculate $\boldsymbol{P}_B^{(i+1)}$, and the difference between $\boldsymbol{P}_B^{(i+1)}$ and $\boldsymbol{P}_B^{(i)}$ is measured. If the difference is not small enough compared to the convergence criteria, $\boldsymbol{R}_H^{(i+1)}$ is calculated using $\boldsymbol{P}_B^{(i+1)}$ for the next iteration, and $\boldsymbol{P}_B$ and $\boldsymbol{R}_H$ are updated with $\boldsymbol{P}_B^{(i+1)}$ and $\boldsymbol{R}_H^{(i+1)}$ respectively. Otherwise, the convergence tag is reversed and $\boldsymbol{P}_B$ and $\boldsymbol{R}_H$ are updated with $\boldsymbol{P}_B^{(i)}$ and $\boldsymbol{R}_H^{(i)}$ respectively.

The convergence speed of the above iterative method is dependent on the parameter setting but experience shows that it is very fast. For example, Table 6.1 lists the numbers of iterations executed to derive results from each model for 10 increasing

session durations. The other parameters are set according to traffic pattern 2, which will be described later. From Table 6.1 it can be observed that the required number of

Model	Numbers of iterations
Random	[2, 2, 3, 4, 5, 7, 9, 11, 11, 13]
RRSS	[2, 2, 3, 4, 5, 7, 8, 11, 12, 13]
WLAN-first	[2, 2, 3, 4, 5, 6, 8, 10, 12, 13]
Service-based	[2, 2, 3, 4, 5, 6, 8, 10, 12, 13]

Table 6.1: *Numbers of iterations executed to derive results from each model for* 10 *increasing session durations listed in Table 6.3*

iterations to reach the convergence criteria for each model are very close. Moreover, since the RAN blocking probabilities get larger when the session duration increases, the iterative method needs more computation to reach the criteria $\epsilon = 10^{-30}$.

6.7 Performance Evaluation

In this work, four strategies are investigated, namely: random, RRSS, WLAN-first and service-based. The effect of different mobility and traffic patterns of a mobile node on the average throughput, RAN blocking probability and handover rate are investigated.

6.7.1 Parameter Settings

Table 6.2 lists the settings of the parameters used in the 2D-CTMC and PEPA models. They are divided up into four groups: (a) the phase branching probabilities of the mobility model which is assumed to have four phases; (b) the data rates of the 3GRAT and WRAT for the NRT and RT sessions[V], and the corrresponding data rate factors; (c) the capacity, coverage percentages of the 3GRAN and the WRAN[VI], and the session arrival rates of the 3G-WLAN interworking cell; (d) the network selection probabilities of different strategies.

The activity rates used in the PEPA models of the random, RRSS, WLAN-first and service-based NSSs are listed in Table 6.3. The duration of the NRT session $1/\mu_{NRT}$

[V] 2 Mbps is used for the data rate of the 3GRAT [126], and 6 Mbps is used for the data rate of the WRAT [125]

[VI] It is assumed that there are 5 circular WLAN cells in a circular 3G-WLAN interworking cell, and their radii are set to 100m and 1000m respectively so that the data rates of 6 Mbps and 2 Mbps can be achieved.

$[a_1, a_2, a_3, a_4]$	$[0.7, 0.5, 0.3, 0]$
$[b_1, b_2, b_3, b_4]$	$[0.3, 0.5, 0.7, 1]$

(a) Branching probabilities

D_{RT}^{C}	2 Mbps	D_{RT}^{W}	2 Mbps
D_{NRT}^{C}	2 Mbps	D_{NRT}^{W}	6 Mbps
R_{NRT}^{C}	1	R_{NRT}^{W}	3

(b) Data rates and data factors

N_C	50	N_W	30
A_C	0.95	A_W	0.05
λ_C^h	1/20	λ_W^h	1/20
Λ^n	1/10		

(c) Capacities, coverage percentages and session arrival rates

Random	$P_C = 0.5$	$P_W = 0.5$
RRSS	$P_C = 0.4$	$P_W = 0.6$ [116]
WLAN-first	$P_C = 0$	$P_W = 1$
Service-based	$P_C = 1$ (for RT session) $P_W = 1$ (for NRT session)	

(d) Network selection probabilities

Table 6.2: *Parameter settings of the 2D-CTMC and PEPA models*

Rate	Descriptions	Value (second^{-1})
λ	session arrival rate	180^{-1}
υ_1, υ_3	mobility rate of phase 1 and phase 3	600^{-1}
υ_2, υ_4	mobility rate of phase 2 and phase 4	$474^{-1}, 1200^{-1}$
μ_{RT}	rate of RT sessions	$[60 : 60 : 600].^{-1}$
μ_{NRT}	rate of NRT sessions	$[30 : 30 : 300].^{-1}$
P_{NRT}	probability of generating a NRT session	0.3, 0.7

Table 6.3: *Activity rates of the PEPA models of the random, RRSS, WLAN-first and service-based strategies*

and that of the RT session $1/\mu_{RT}$ are the control parameters in all the evaluation. The duration of the NRT session is measured at 2 Mbps and thus it reflects the traffic volume of the NRT session. The session arrival interval of a mobile node is set to 180 seconds. The mean sojourn time of the mobile node in the 3G area ($1/\upsilon_1$ and $1/\upsilon_3$) is assumed to be the same and is set to 600 seconds. Two mobility patterns and two traffic patterns are investigated, and they are controlled by the mean sojourn time of the mobile node in the 3G-WLAN area ($1/\upsilon_2$ and $1/\upsilon_4$) and the proportion of the NRT session generated by the mobile node (P_{NRT}).

6.7.2 Effect of Mobility Pattern

The mobility pattern of the mobile node is controlled by its mean sojourn time in the 3G-WLAN area. Two patterns are considered in the evaluation.

1. In the first pattern, the fluid flow movement model [127] is employed. In this model, the mean sojourn time of a mobile node in an area is calculated as:

$$t_{sojourn} = \frac{\pi A}{\bar{\eta} L}, \tag{6.21}$$

where $\bar{\eta}$ is the average speed of the mobile node, L and A are the perimeter and area of the region with arbitrary shape. Since the WLAN cells are assumed to not overlap with each other, the sojourn time in the 3G area is:

$$t_{3G} = \frac{\pi(\pi * 1000^2 - 5 * \pi * 100^2)}{\bar{\eta}(2\pi * 1000 + 5 * 2\pi * 100)}, \tag{6.22}$$

and the sojourn time in the 3G-WLAN area is:

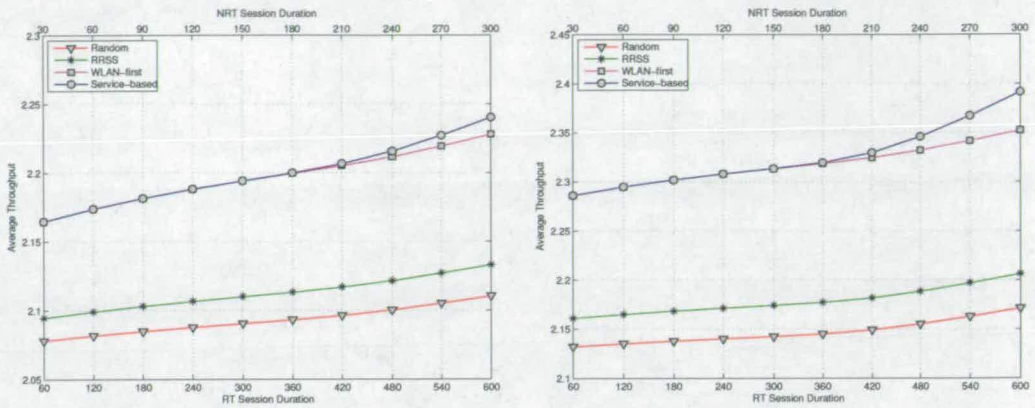
$$t_{3G\text{-}WLAN} = 5 * \frac{\pi * (\pi * 100^2)}{\bar{\eta} * (2\pi * 100)}. \tag{6.23}$$

Hence, the ratio of t_{3G} and $t_{3G\text{-}WLAN}$ is 38/30, and if t_{3G} is 600 seconds then $t_{3G\text{-}WLAN}$ is about 474 seconds.

2. In the second pattern, the mobile node spends a longer time in the 3G-WLAN area and υ_2 and υ_4 are set to 1/1200.

Note that the number of phases, the sojourn times and the branching probabilities of the mobility model used in the evaluation are contrived, more practical values can be estimated from field data using algorithms studied in [122, 124]. The mobile node generates the NRT and RT sessions with equal probabilities, i.e., $P_{NRT} = P_{RT} = 0.5$. The other parameters and activity rates are the same as listed in Table 6.2 and Table 6.3. In all the figures, the investigated performance measures are plotted against the session duration. Since there are two types of sessions, two x-axes are used where the top x-axis is the NRT session duration and the bottom x-axis is the RT session duration.

6.7.2.1 Average Throughput



(a) Mobility pattern 1 ($t_{3G\text{-}WLAN} = 474$, $P_{NRT} = P_{RT} = 0.5$)

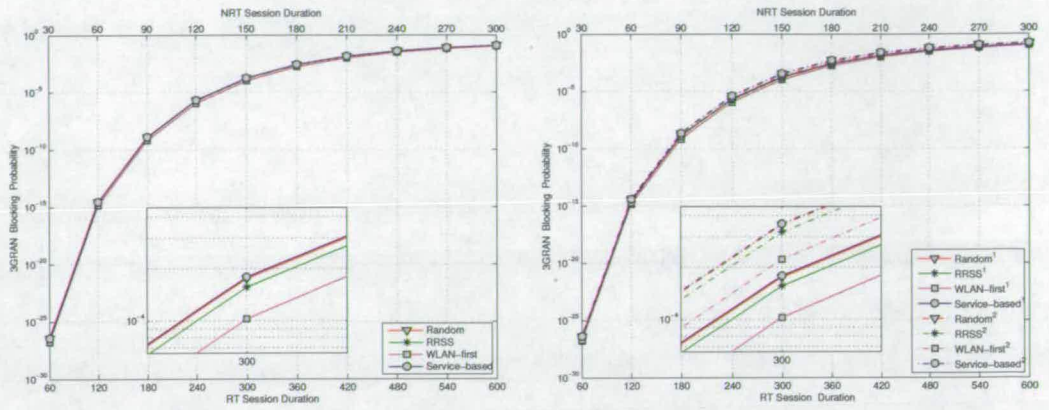
(b) Mobility pattern 2 ($t_{3G\text{-}WLAN} = 1200$, $P_{NRT} = P_{RT} = 0.5$)

Figure 6.7: *The effect of mobility pattern on average throughput*

Figure 6.7 shows the average throughput achieved by the mobile node with different mobility patterns. According to Eq. (6.6), if $D_W^{NRT} = 2$, then the average throughput is always 2 Mbps. Therefore, the average throughput mostly depends on how much time an NRT session uses the WRAT. Figure 6.7 indicates that a higher average throughput can be gained by using the strategy with a larger WRAN selection probability, or by simply staying in the 3G-WLAN area for a longer time. Moreover, a longer NRT session duration also results in a higher average throughput, because the mobile node has more opportunity to use the WRAT. An interesting observation is that when NRT session duration is less than 210 seconds, the WLAN-first and service-based strategies have almost the same performance on average throughput but for longer session durations, more of an NRT session in the service-based strategy is spent on the WRAT which results in a clear improvement on the average throughput. This result suggests that the service-based strategy can make the best use of the high data rate of the WRAN, especially for long NRT sessions.

6.7.2.2 RAN Blocking Probability

Figure 6.8 shows the blocking probability of the 3GRAN experienced by the mobile node with different mobility patterns. Note that the y-axis is a logarithmic scale. Since the 2D-CTMC resource consumption model does not include any handover prioritised scheme, the derived blocking probability is for both new and handover



(a) Mobility pattern 1 ($t_{3G\text{-}WLAN} = 474$, $P_{NRT} = P_{RT} = 0.5$)

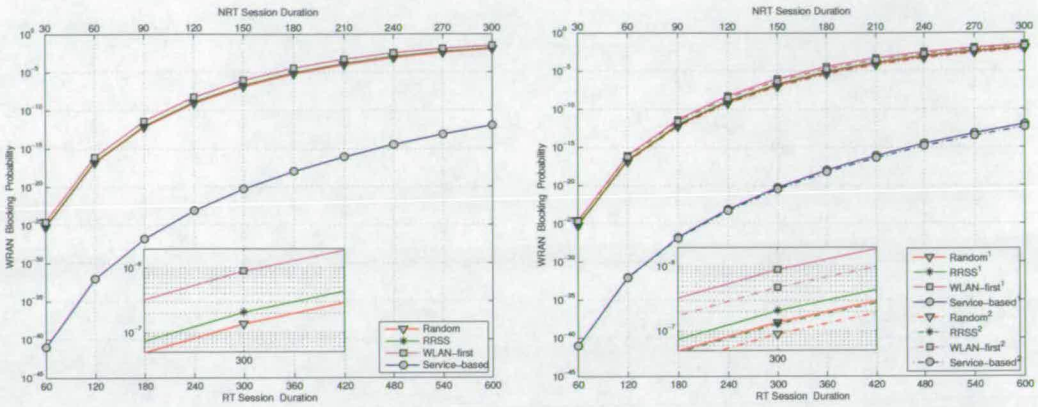
(b) Mobility pattern 1 ($t_{3G\text{-}WLAN} = 474$, $P_{NRT} = P_{RT} = 0.5$) and Mobility pattern 2 ($t_{3G\text{-}WLAN} = 1200$, $P_{NRT} = P_{RT} = 0.5$)

Figure 6.8: *The effect of mobility pattern on 3GRAN blocking probability*

session requests. The results indicate that the blocking probability of the 3GRAN mostly depends on its traffic load, which is decided by how frequently it is chosen for a session and how long the session engages the resources. From Figure 6.8(a), it can be observed that the service-based and random strategies are very close and also have the highest blocking probabilities. This is because given $P_{NRT} = P_{RT} = 0.5$, the random and service-based strategies have the same and also the highest probability of using 3GRAN. This can be explained as follows: in the service-based strategy, although only RT sessions choose the 3GRAN, the probability that a session is RT is 0.5; whereas in the random strategy, both types of sessions can choose the 3GRAN with the probability of 0.5. However, since the approximation made in Eq. (6.13) is an underestimate[VII], the service-based strategy should have a clearly higher 3GRAN blocking probability than the random strategy. The RRSS strategy ranks third as all the sessions choose the 3GRAN with the probability of 0.4 and the WLAN-first strategy has the lowest 3GRAN blocking probability. In Figure 6.8(b), the effect of the mobile node's mobility is shown. For all of the strategies, a longer stay in the 3G-WLAN area results in a higher 3GRAN blocking probability.

Similarly, the blocking probabilities of the WRAN also depends on its traffic load and the differences between the different strategies are more obvious as shown in Figure 6.9. It can be observed that the WRAN blocking probability of the WLAN-

[VII]This is because the RT session, which has a longer duration, should have a proportion larger than P_{RT}.



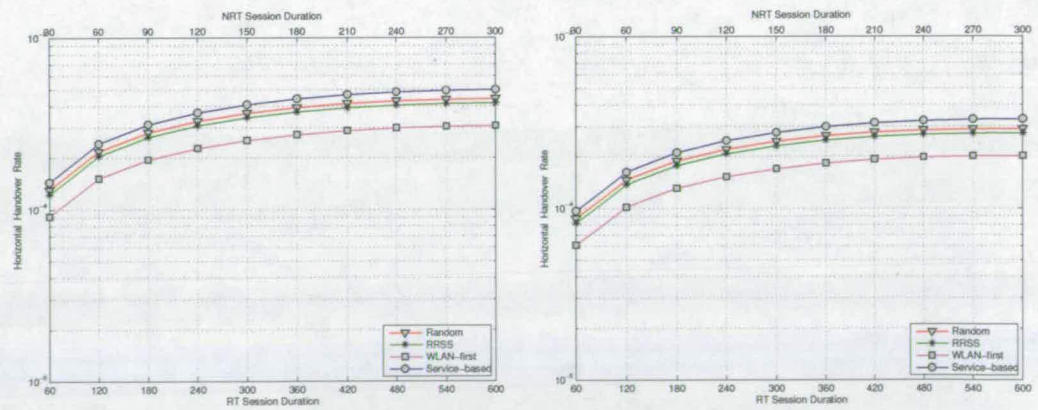
(a) Mobility pattern 1 ($t_{3G\text{-}WLAN} = 474$, $P_{NRT} = P_{RT} = 0.5$)

(b) Mobility pattern 1 ($t_{3G\text{-}WLAN} = 474$, $P_{NRT} = P_{RT} = 0.5$) and Mobility pattern 2 ($t_{3G\text{-}WLAN} = 1200$, $P_{NRT} = P_{RT} = 0.5$)

Figure 6.9: *The effect of mobility pattern on WRAN blocking probability*

first strategy is the largest, as expected, and that of the service-based strategy is much smaller than the others because only NRT sessions are allowed to use the WRAT and more importantly they engage the WRAN resources for a shorter time than the RT sessions. Moreover, the mobility of the mobile node has the opposite effect on the WRAN blocking probability to that on the 3GRAN, i.e., a longer stay in the 3G-WLAN area results in a lower WRAN blocking probability.

6.7.2.3 Handover Rate



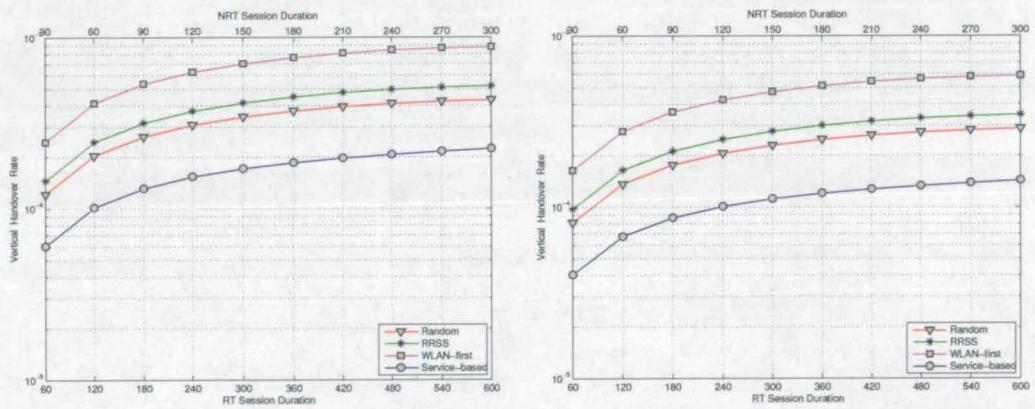
(a) Mobility pattern 1 ($t_{3G\text{-}WLAN} = 474$, $P_{NRT} = P_{RT} = 0.5$)

(b) Mobility pattern 2 ($t_{3G\text{-}WLAN} = 1200$, $P_{NRT} = P_{RT} = 0.5$)

Figure 6.10: *The effect of mobility pattern on horizontal handover rate*

Figure 6.10 shows the horizontal handover rate performed by the mobile node with

different mobility patterns. A horizontal handover happens when the engaged mobile node moves across adjacent 3G-WLAN interworking cells. According to Eq. (6.17), the horizontal handover rate depends on the probability that the mobile node is using the 3GRAT at the time it moves out of its current 3G-WLAN interworking cell. The results indicate that a mobile node using the service-based strategy is the most likely to perform a horizontal handover. This is because the service-based strategy makes the mobile node spend longer in the 3GRAN than the other strategies as explained as explained in Section 6.7.2.2. The random strategy ranks the second and the WLAN-first strategy has the smallest horizontal handover rate as it uses the 3GRAT less than the other strategies. Given a certain mobility pattern, a longer session duration means the mobile node is more likely to hand over during a session and thus results in a higher handover rate. Moreover, given a certain session duration, lower mobility can reduce the handover rate for all of the strategies, as shown in Figure 6.10(b).



(a) Mobility pattern 1 ($t_{3G\text{-}WLAN} = 474$, $P_{NRT} = P_{RT} = 0.5$)

(b) Mobility pattern 2 ($t_{3G\text{-}WLAN} = 1200$, $P_{NRT} = P_{RT} = 0.5$)

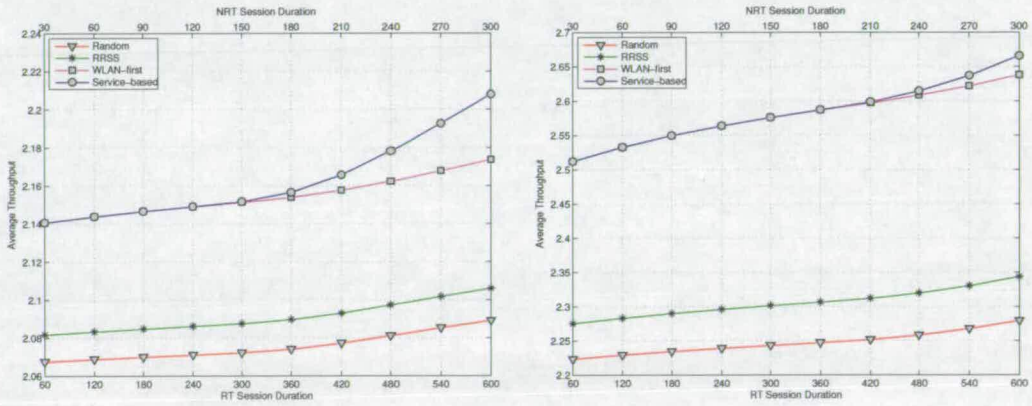
Figure 6.11: *The effect of mobility pattern on vertical handover rate*

The vertical handover is defined as the handover between different RATs. This rate depends on the probability that the mobile node is using the WRAT at the time it moves out of the 3G-WLAN area (Eq. (6.18) and Eq. (6.20)), and the probability of choosing the WRAN when it moves into the 3G-WLAN area (Eq. (6.19)). Therefore, as shown in Figure 6.11, the WLAN-first strategy experiences the most frequent vertical handover whereas the service-based strategy has the lowest vertical handover rate. The effect of the session duration and mobility on the vertical handover rate are the same as those on the horizontal handover rate.

6.7.3 Effect of Traffic Pattern

Two traffic patterns are considered in the evaluation and they are controlled by how frequently the mobile node generates RT and NRT sessions. In the first and second pattern, P_{NRT} is set to 0.3 and 0.7 respectively. All the other parameters are as defined in Table 6.2 and Table 6.3. The slow mobility pattern is used, that is, the sojourn time of the mobile node in the 3G-WLAN area is set to 1200 seconds. Again, two x-axes are used in all the figures to show the durations of different types of sessions.

6.7.3.1 Average Throughput



(a) Traffic pattern 1 ($P_{NRT} = 0.3$, $t_{3G\text{-}WLAN} = 1200$)

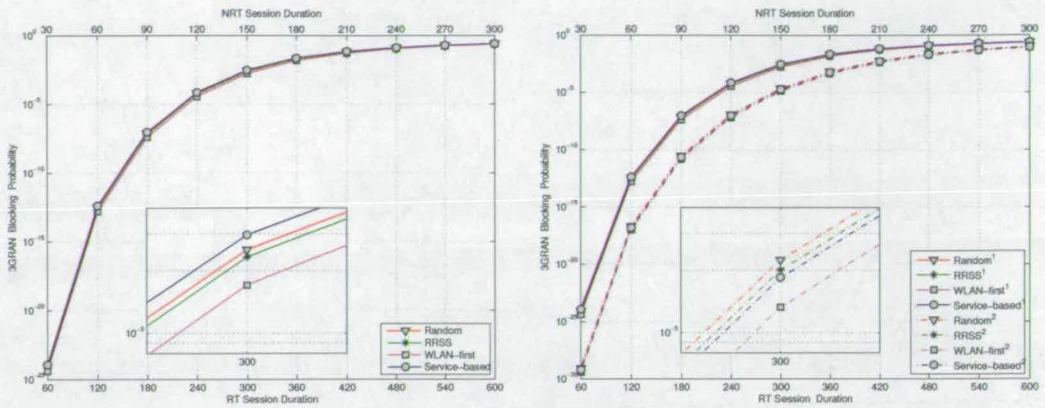
(b) Traffic pattern 2 ($P_{NRT} = 0.7$, $t_{3G\text{-}WLAN} = 1200$)

Figure 6.12: *The effect of traffic pattern on average throughput*

Similar trends for the average throughput of different strategies as in Section 6.7.2.1 can be observed in Figure 6.12. Moreover, by comparing Figure 6.7(b) and Figure 6.12, it can be found that a higher NRT probability results in a larger average throughput since in this case there will be more NRT sessions that use the WRAT.

6.7.3.2 RAN Blocking Probability

Figure 6.13 shows the 3GRAN blocking probability experienced by the mobile node with different traffic patterns. As discussed in Section 6.7.2.2, the blocking probability of a certain RAN mostly depends on how frequently it is chosen for a session and how long the session engages the resources. In Figure 6.13(a) where $P_{NRT} = 0.3$, the service-based strategy has the heaviest traffic load since 70% of the traffic are RT sessions which choose the 3GRAN. The random strategy comes second with all the



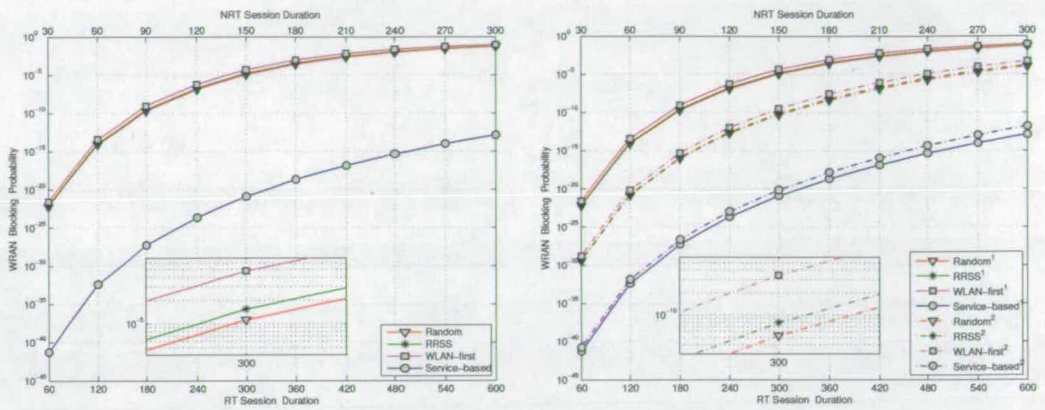
(a) Traffic pattern 1 ($P_{NRT} = 0.3$, $t_{3G\text{-}WLAN} = 1200$)

(b) Traffic pattern 1 ($P_{NRT} = 0.3$, $t_{3G\text{-}WLAN} = 1200$) and Traffic pattern 2 ($P_{NRT} = 0.7$, $t_{3G\text{-}WLAN} = 1200$)

Figure 6.13: *The effect of traffic pattern on 3GRAN blocking probability*

sessions choosing the 3GRAN with a probability of 0.5, followed by the RRSS strategy which has a 3GRAN selection probability of 0.4 and the WLAN-first strategy has the lowest 3GRAN blocking probability. In the second traffic pattern where $P_{NRT} = 0.7$, the 3GRAN blocking probabilities of all the strategies are reduced as shown in Figure 6.13(b), which is mainly because in the parameter settings the duration of an NRT session is shorter than that of an RT session. Therefore, a higher percentage of NRT sessions reduces the resource engagement time of the 3GRAN. Moreover, the 3GRAN blocking probability of the service-based strategy is very sensitive to the traffic pattern and is reduced by a larger amount than the others and is lower than those of the random and RRSS strategies.

The advantage of the service-based strategy on the WRAN blocking probability is very clear as shown in Figure 6.14(a) where a larger probability of choosing the WRAN results in a higher WRAN blocking probability. An interesting observation is that at a higher NRT probability, the WRAN blocking probability of the service-based strategy grows whereas those of the other strategies are reduced as shown in Figure 6.14(b). The reason is that in the service-based strategy the WRAN resources are only engaged by the NRT sessions whereas in the other strategies the WRAN resources can be engaged by all types of sessions. As a result, a higher NRT probability implies a higher WRAN traffic load in the service-based strategy, whereas in the other strategies this means there will be fewer RT sessions that cannot make use of the high data rate of WRAT and thus engage the WRAN resources.

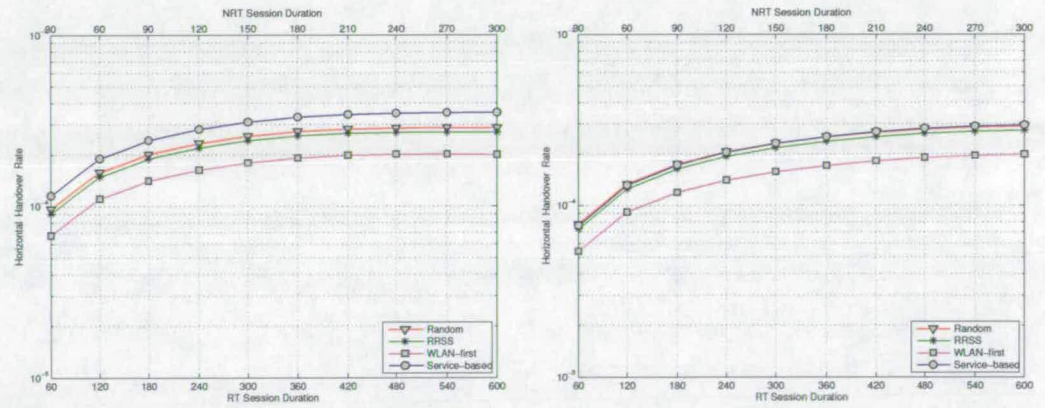


(a) Traffic pattern 1 ($P_{NRT} = 0.3$, $t_{3G\text{-}WLAN} = 1200$)

(b) Traffic pattern 1 ($P_{NRT} = 0.3$, $t_{3G\text{-}WLAN} = 1200$) and Traffic pattern 2 ($P_{NRT} = 0.7$, $t_{3G\text{-}WLAN} = 1200$)

Figure 6.14: *The effect of traffic pattern on WRAN blocking probability*

6.7.3.3 Handover Rate



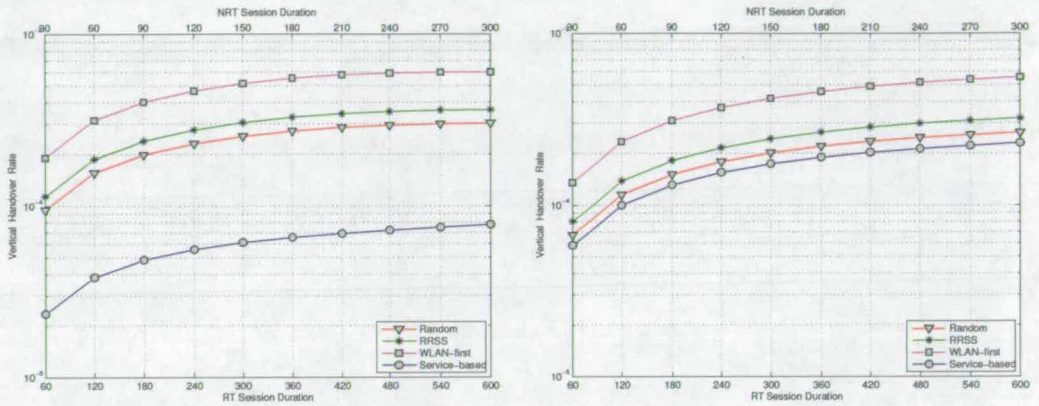
(a) Traffic pattern 1 ($P_{NRT} = 0.3$, $t_{3G\text{-}WLAN} = 1200$)

(b) Traffic pattern 2 ($P_{NRT} = 0.7$, $t_{3G\text{-}WLAN} = 1200$)

Figure 6.15: *The effect of traffic pattern on horizontal handover rate*

The effect of traffic pattern on the horizontal handover rate is shown in Figure 6.15. As can be observed by comparing Figure 6.10(b) and Figure 6.15, the traffic pattern changes the horizontal handover rate of the random, RRSS and WLAN-first strategies to a small extent since their selection strategies are not based on the type of the session. As explained in Section 6.7.3.2, a higher percentage of NRT sessions reduces the time the mobile node is connected to the 3GRAN. Therefore, the horizontal handover rates of the these strategies are reduced. On the other hand, since the service-based strategy

only allows RT sessions to use the 3GRAT, a lower RT probability means that a mobile node is less likely to be connected to the 3GRAN and thus has a smaller horizontal handover rate. This is why when $P_{NRT} = 0.7$, the horizontal handover rate of the service-based strategy is reduced and is almost the same as that of the random strategy.



(a) Traffic pattern 1 ($P_{NRT} = 0.3$, $t_{3G\text{-}WLAN} = 1200$)

(b) Traffic pattern 2 ($P_{NRT} = 0.7$, $t_{3G\text{-}WLAN} = 1200$)

Figure 6.16: *The effect of traffic pattern on vertical handover rate*

A similar effect of the traffic pattern on the vertical handover rate can be observed in Figure 6.16. A larger NRT probability results in a higher vertical handover rate in the service-based strategy since the mobile node uses the WRAT more frequently and thus is more likely to be connected to the WRAN. Unlike the horizontal handover rate, the vertical handover rate in the other strategies is more sensitive to the traffic pattern; they decrease as the time the mobile node is connected to the WRAN is reduced at larger NRT probability.

6.8 Conclusions

To find out the effect of different NSSs on the performance of both mobile nodes and RANs, in this chapter a general performance evaluation framework for NSSs is investigated. This framework is general because it has an interface to the NSS used by the mobile node and an interface to the resource consumption model of the RANs. Four types of strategies, namely random, RRSS, WLAN-first and service-based, have been evaluated from different perspectives.

The three types of performance measures discussed in this work are meaningful from both the user and the network administrator's perspectives. Both average

throughput and handover rate have effect on the QoS perceived by the user. The average throughput is important as it reflects the efficiency of the communication especially for NRT sessions. The handover rate indicates the volume of signalling load and the frequency of service interruption during a session. Therefore a high handover rate should be avoided and in particular vertical handovers, since their cost is higher than that of horizontal handovers due to more involved process. On the other hand, the network administrator may be more concerned about resource utilisation of RANs and the RAN blocking probability can reflect the traffic loads of different RANs.

The deterministic strategies, such as the WLAN-first and service-based strategies, are easy to implement and a user always knows which RAN is going to be selected. Since the WLAN-first strategy chooses the WLAN whenever it is available, it has the lowest 3GRAN blocking probability and horizontal handover rate, at the expense of having the highest WRAN blocking probability and vertical handover rate. It can also achieve high average throughput but is outperformed by the service-based strategy at long session durations. On the other hand, the service-based strategy makes the best use of the high data rate of the WRAT by only allowing NRT sessions to access the WRAN and consequently has the lowest WRAN blocking probability. Since the service-based strategy is aware of the type of the session, its performance is very sensitive to the traffic pattern of the mobile node. That is, the RT probability is proportional to the 3GRAN blocking probability and horizontal handover rate, and is inversely proportional to the vertical handover rate. As for the non-deterministic strategies, the random and RRSS strategies introduce randomness in network selection and therefore the user will experience uncertainty during the handover. As can be seen from the results, they have more balanced performance on the investigated measures than the deterministic strategies. This phenomenon is likely to extend to other non-deterministic strategies as well since they have intermediate probabilities of choosing the WRAN and 3GRAN.

The effect of the mobility pattern of the mobile node is straightforward. A longer sojourn time in the 3G-WLAN area results in a higher average throughput, and lower horizontal and vertical handover rates. An interesting observation is that the mobile node will experience a higher 3GRAN blocking probability and a lower WRAN blocking probability if the sojourn time in the 3G-WLAN area is longer. As for the effect of the traffic pattern, the attribute of the service-based strategy means it is

strongly affected by the traffic pattern, whereas the other strategies are affected simply because the session durations are different when traffic pattern changes.

Chapter 7
Conclusions

This final chapter draws conclusions from the results and contributions presented in the previous chapters. The limitations of this work and several avenues for future work are also considered.

7.1 Conclusions

The work presented in this thesis addresses performance modelling of network management schemes for mobile wireless networks, using a formal performance modelling formalism named PEPA. The modelled network management schemes have been designed with the aim of providing seamless and high-quality services to users in mobile environments. These schemes achieve the objective by administering network resources and regulating behaviour of mobile nodes so that performance measures that determine the QoS perceived by users can be improved. In essence, mobile wireless networks are resource-sharing systems, and management schemes for such systems specify how component parts of mobile wireless networks, such as users, protocols and network entities, interact with each other. As has been demonstrated by the PEPA models in the previous chapters, the PEPA language provides a great deal of flexibility in model construction since its inherent compositionality and concurrency enable the expression of how network components in mobile wireless networks are structured and also how these components cooperate with each other as specified by the modelled network management schemes. Performance evaluation based on a PEPA model is straightforward in that there is a corresponding CTMC underlying the PEPA model and performance measures can be derived from both steady state and transient analysis of the model (only steady state analysis is carried out in this thesis though). All the PEPA models in this thesis are constructed in such a way that they have clear and accurate representations of the mechanisms underlying the modelled schemes, explaining how the schemes meet their design objectives. Model generality is also maintained so that the models are independent of detailed

implementations. In brief, this thesis has practically demonstrated that the PEPA language is an ideal modelling technique for an initial and indispensable investigation of system performance, before more complex performance models or simulation considering system implementation details are developed.

Two important issues regarding seamless and high-quality service provisioning in mobile wireless networks are investigated in this thesis. The first issue is the deployment, in mobile environments, of RSVP which was designed to achieve end-to-end and guaranteed QoS in wired networks. Since RSVP reserves resources on the basis of a flow which is identified by the IP addresses of communicating ends, a mobile node has to request a new reservation path after a network layer handover in order to continue its QoS session. There are two major problems associated with this reservation re-establishment process. First of all, the old and new reservation paths between the mobile node and its correspondent node before and after a handover are usually overlapped in part. Therefore, it is desirable that the RSVP signalling should be localised within the affected part of the network. This problem is discussed in Chapter 4, in which the basic and mobility-supported RSVP are modelled and evaluated. In the PEPA models, networks are separated into two parts; network resources are abstracted as channels; mobile nodes consume resources according to the characteristics of these schemes. The performance of these schemes are compared in terms of the blocking probability of reservation requests and the RSVP signalling cost after a handover. The results indicate that the mobility-supported RSVP clearly outperforms the basic RSVP on both measures, and that the former is more suitable in high traffic and mobility scenarios. The evaluation also highlights that these enhancements are achieved by the optimisation of the basic RSVP signalling procedure, avoiding unnecessary resource reservation paths in the unaffected part of the network and limiting RSVP signalling within the affected part.

With only the signalling optimisation scheme, interruptions can still happen to mobile nodes if there are not enough resources at the mobile nodes' new locations. For that reason, another major problem of deploying RSVP in mobile wireless networks is the design of ARR schemes that reserve resources in advance for the mobile nodes. Since conventional ARR schemes do not discriminate between requests that actively or passively reserve resources, they waste too many network resources from the QoS traffic's perspective. To solve this problem, a reservation optimised ARR scheme

is proposed in Chapter 5. The proposed ARR scheme consists of two admission control mechanisms, i.e. passive reservation limited and SMR-based replacement, that can be easily integrated into existing ARR schemes, and consequently maintains its modularity. These mechanisms aim to restrict the number of advance reservation paths in a network that are not actively used by the mobile nodes, and to take account of the traffic and mobility patterns of the mobile nodes, only allowing the most eligible ones to reserve resources in advance. Models of different types of ARR schemes are constructed and assessed to demonstrate the advantages of the proposed scheme, from the perspectives of the blocking probabilities of active and passive reservation requests and the mean numbers of active and passive reservation paths in a network. Results of the investigation on these performance measures indicate that the proposed scheme, by setting aside dedicated resources for actively used reservation paths and restricting the ability of slow mobile nodes to make advance reservations, attains a better network resource utilisation. The improvements of the proposed scheme are gained at the expense of introducing possible handover interruptions to the slow mobile nodes, which has been argued to be reasonable.

In Chapter 4 and Chapter 5, no assumption is made on the access technologies used in the mobile wireless networks and they are transparent to the studied network management schemes. However, the main trend for future wireless communications is a shift to hybrids of different types of wireless networks. Therefore, the second issue studied in this thesis considers a heterogeneous wireless network environment. Mobile users in this heterogeneous environment may not be satisfied with just reliable connectivity, and rather they may also want to access their services through the best possible networks, considering their preferences, application demands, available network resources, etc. To meet this requirement, NSSs have been designed to abstract the process of selecting a suitable network as mathematical expressions which assess alternative networks in terms of different criteria. In Chapter 6, a general performance evaluation framework for the NSSs used in a particularly popular heterogeneous environment, i.e. 3G-WLAN interworking networks, is investigated. This framework captures the multi-service feature of the next-generation wireless communications in its traffic model and characterises various tracks of a mobile node within and across 3G-WLAN interworking cells in its mobility model. The framework retains its generality by having an interface to the NSSs used by mobile nodes and an

interface to the model capturing how resources of the RANs are consumed. These two interfaces are abstracted in the form of the network selection probability and the RAN blocking probability respectively. This approach provides the framework with flexibility since they can be used as independent input parameters. Moreover, a novel iterative algorithm is also proposed to derive RAN blocking probability from this framework when it is not at the modellers' disposal. This algorithm links models of network resource consumption and PEPA models of different NSSs by interchanging necessary parameters between them. The convergence speed of the algorithm is fast, and as well as evaluating RAN blocking probability, handover rate is also determined automatically. Deterministic strategies that choose networks according to some on-line measures and non-deterministic strategies that select certain networks as specified are studied. Performance measures that are meaningful from the user's perspective (i.e. average throughput and handover rate) and the network administrator's perspective (i.e. RAN blocking probability) are evaluated. The assessment of these measures explore the effect of the studied strategies on the communication efficiency and service quality perceived by the mobile nodes and on the traffic loads of access networks. Their characteristics in different mobility and traffic patterns are also presented. Each type of strategy has its own advantages and disadvantages and the usage of them depends on the user and network requirements.

7.2 Limitations of This Thesis and Suggestions for Future Work

There are also limitations that need to be acknowledged and addressed regarding this thesis. First of all, although the PEPA models built in this thesis have no restriction on model size, small-size models are used in evaluation in order that they are numerically tractable. Therefore, model simplification, such as careful design of models and equivalence reasoning based on the PEPA language, should be applied so that evaluation can be carried out on larger models. Another limitation is that the PEPA models are built on simple statistical assumptions and are independent of the implementation of the studied schemes. Therefore, evaluation results from the performance models have not been compared to field data collected from experiments or simulation which employ more sophisticated probability distribution functions. The results of different techniques can be different but they should share the same

trend. Such a comparison is conducted in [128], in which PEPA models are constructed to identify the impacts of bottlenecks of a network. These models generate results that are comparable to those derived from multicommodity flow analysis and draw more general rules about deploying bottlenecks in the network.

Some areas of future research that could be pursued based on this thesis can be identified:

- Although the integration of HMIPv6 and RSVP has been demonstrated to be an easy solution to RSVP signalling optimisation in Chapter 4, reservation paths could still overlap in part within the MAP domain. Accordingly, approaches that can efficiently identify common paths should be developed.

- In the proposed reservation optimised ARR scheme in Chapter 5, QoS sessions of mobile nodes are assumed to be of the same type and only those with large SMR values are eligible to be allocated passive reservation paths. However, QoS class is a very important parameter to determine which mobile node is most appropriate for making reservations in advance. Performance models capturing this feature should be investigated.

- With RSVP signalling optimisation and ARR schemes, service disruptions to mobile nodes during handover can be reduced. However, the lack of careful design of the signalling protocols of these schemes will hinder their deployment. This design includes details such as what kind of information should be carried in signalling messages, which entities in a network should process these messages, and how to reduce signalling processing overheads, etc. Benefits of the carefully designed protocols may also be demonstrated through performance modelling.

- In the network resource consumption model in Chapter 6, it is assumed that no CAC algorithm is used. In fact, various CAC algorithms can be employed by RANs in HMWNs. The simplest example is the algorithm in which handover sessions are given high priority over new sessions. Different algorithms should be considered in the evaluation of NSSs, and by using the proposed performance evaluation framework, optimum algorithms which benefit both users and network administrators may be designed.

- Moreover, in the work presented in Chapter 6, all mobile users are assumed to adopt the same NSS. However, this is rarely the case in reality. It would be very

useful if a single performance model which expresses a population of mobile users using different NSSs could be constructed. From this model, the effect of user behaviour on the RAN will be identified.

References

[1] A. R. Prasad and N. R. Prasad, *802.11 WLANs and IP networking: security, QoS, and mobility*. Artech House, 2005.

[2] Bluetooth. [Online]. Available: https://www.bluetooth.org/apps/content/

[3] *Wireless medium access control (MAC) and physical layer (PHY) specifications for wireless personal area networks (WPANs)*, IEEE Std. 802.15.1, 2005.

[4] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications*, IEEE Std. 802.11, 1999.

[5] R. Lloyd-Evans, *QoS in Integrated 3G Networks*. Artech House, 2002.

[6] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Prentice Hall, 2002.

[7] E. Gustafsson and A. Jonsson, "Always best connected," *IEEE Wireless Commun. Mag.*, vol. 10, no. 1, pp. 49–55, 2003.

[8] C. Perkins, "IP Mobility Support for IPv4," IETF RFC 3344, Aug. 2002.

[9] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6," IETF RFC 3775, Jun. 2004.

[10] H. Soliman, C. Castelluccia, K. E. Malki, and L. Bellier, "Hierarchical Mobile IPv6 (HMIPv6) Mobility Management," IETF RFC 5380, Oct. 2008.

[11] C. Perkins, "Mobile IP," *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 66–82, 2002.

[12] ——, "IP Encapsulation within IP," IETF RFC 2003, Oct. 1996.

[13] C. Perkins and D. Johnson, "Route Optimization in Mobile IP," draft-ietf-mobileip-optim-11.txt, Sep. 2001.

[14] A. Conta and S. Deering, "Generic Packet Tunneling in IPv6 Specification," IETF RFC 2473, Dec. 1998.

[15] R. Koodli and C. Perkins, "Mobile IPv4 Fast Handovers," IETF RFC 4988, Oct. 2007.

[16] R. Koodli, "Mobile IPv6 Fast Handovers," IETF RFC 5268, Jun. 2008.

[17] A. Campbell and J. Gomez-Castellanos, "IP micro-mobility protocols," *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 4, no. 4, pp. 45–53, 2000.

[18] P. Reinbold and O. Bonaventure, "IP micro-mobility protocols," *IEEE Commun. Surveys Tuts.*, vol. 5, no. 1, pp. 40–57, 2003.

[19] J. Gozdecki, A. Jajszczyk, and R. Stankiewicz, "Quality of service terminology in IP networks," *IEEE Commun. Mag.*, vol. 41, no. 3, pp. 153–159, 2003.

[20] R. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: an Overview," IETF RFC 1633, Jun. 1994.

[21] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Service," IETF RFC 2475, Dec. 1998.

[22] S. Shenker, C. Partridge, and R. Guerin, "Specification of Guaranteed Quality of Service," IETF RFC 2212, Sep. 1997.

[23] J. Wroclawski, "Specification of the Controlled-Load Network Element Service," IETF RFC 2211, Sep. 1997.

[24] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification," IETF RFC 2205, Sep. 1997.

[25] A. Mankin, F. Baker, B. Braden, S. Bradner, M. O'Dell, A. Romanow, A. Weinrib, and L. Zhang, "Resource ReSerVation Protocol (RSVP) – Version 1 Applicability Statement Some Guidelines on Deployment," IETF RFC 2208, Sep. 1997.

[26] Y. Bernet, "The complementary roles of RSVP and differentiated services in the full-service QoS network," *IEEE Commun. Mag.*, vol. 38, no. 2, pp. 154–162, 2000.

[27] V. Jacobson, K. Nichols, and K. Poduri, "An Expedited Forwarding PHB," IETF RFC 2598, Jun. 1999.

[28] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, "Assured Forwarding PHB Group," IETF RFC 2597, Jun. 1999.

[29] S. Giordano, S. Salsano, S. Berghe, G. Ventre, and D. Giannakopoulos, "Advanced QoS provisioning in IP networks: the European premium IP projects," *IEEE Commun. Mag.*, vol. 41, no. 1, pp. 30–36, 2003.

[30] Y. Bernet, P. Ford, R. Yavatkar, F. Baker, L. Zhang, M. Speer, R. Braden, B. Davie, J. Wroclawski, and E. Felstaine, "A Framework for Integrated Services Operation over Diffserv Networks," IETF RFC 2998, Nov. 2000.

[31] B. Moon and H. Aghvami, "RSVP extensions for real-time services in wireless mobile networks," *IEEE Commun. Mag.*, vol. 39, no. 12, pp. 52–59, 2001.

[32] ——, "Diffserv extensions for QoS provisioning in IP mobility environments," *IEEE Wireless Commun. Mag.*, vol. 10, no. 5, pp. 38–44, 2003.

[33] C. Makaya and S. Pierre, "An architecture for seamless mobility support in IP-Based next-generation wireless networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 2, pp. 1209–1225, 2008.

[34] S. Y. Hui and K. H. Yeung, "Challenges in the migration to 4G mobile systems," *IEEE Commun. Mag.*, vol. 41, no. 12, pp. 54–59, 2003.

[35] M. Kassar, B. Kervella, and G. Pujolle, "An overview of vertical handover decision strategies in heterogeneous wireless networks," *Elseview Comput. Commun.*, vol. 31, no. 10, pp. 2607–2620, 2008.

[36] M. Stemm and R. H. Katz, "Vertical handoffs in wireless overlay networks," *Springer Mob. Netw. Appl.*, vol. 3, no. 4, pp. 335–350, 1998.

[37] J. McNair and F. Zhu, "Vertical handoffs in fourth-generation multinetwork environments," *IEEE Wireless Commun. Mag.*, vol. 11, no. 3, pp. 8–15, 2004.

[38] N. Nasser, A. Hasswa, and H. Hassanein, "Handoffs in fourth generation heterogeneous networks," *IEEE Commun. Mag.*, vol. 44, no. 10, pp. 96–103, 2006.

[39] H. Wang, R. Katz, and J. Giese, "Policy-enabled handoffs across heterogeneous wireless networks," in *Proc. IEEE International Workshop on Mobile Computing Systems and Applications '99*, 1999, pp. 51–60.

[40] A. Hasswa, N. Nasser, and H. Hossanein, "Generic vertical handoff decision function for heterogeneous wireless networks," in *Proc. IFIP International Conference on Wireless and Optical Communications Networks '05*, 2005, pp. 239–243.

[41] F. Bari and V. Leung, "Automated network selection in a heterogeneous wireless network environment," *IEEE Netw.*, vol. 21, no. 1, pp. 34–40, 2007.

[42] W. Zhang, "Handover decision using fuzzy MADM in heterogeneous networks," in *Proc. IEEE Wireless Communications and Networking Conference '04*, 2004, pp. 653–658.

[43] A. Calvagna and G. Modica, "A user-centric analysis of vertical handovers," in *Proc. ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots '04*, 2004, pp. 137–146.

[44] I. Joe, W.-T. Kim, and S. Hong, "A network selection algorithm considering power consumption in hybrid wireless networks," in *Proc. IEEE International Conference on Computer Communications and Networks '07*, 2007, pp. 1240–1243.

[45] K. Yang, I. Gondal, and B. Qiu, "Multi-dimensional adaptive SINR based vertical handoff for heterogeneous wireless networks," *IEEE Commun. Lett.*, vol. 12, no. 6, pp. 438–440, 2008.

[46] E. Stevens-Navarro and V. Wong, "Comparison between vertical handoff decision algorithms for heterogeneous wireless networks," in *Proc. IEEE Vehicular Technology Conference-Spring '06*, 2006, pp. 947–951.

[47] J.-Z. Sun, "A review of vertical handoff algorithms for cross-domain mobility," in *Proc. IEEE International Conference on Wireless Communications and Networking '07*, 2007, pp. 3156–3159.

[48] H. Kobayashi, *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology*, ser. The Systems Programming Series. Addison-Wesley Publishing Company, 1978.

[49] M. C. Jeruchim, P. Balaban, and K. S. Shanmugan, *Simulation of Communication Systems: Modeling, Methodology and Techniques*. Kluwer Academic Publishers, 2002.

[50] P. G. Harrison and N. M. Patel, *Performance Modelling of Communication Networks and Computer Architectures*. Addison-Wesley Publishing Company, 1992.

[51] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. John Wiley & Sons, 1998.

[52] U. Herzog, "Formal Methods for Performance Evaluation," in *Lectures on Formal Methods and Performance Analysis*, ser. Lecture Notes in Computer Science. Springer-Verlag, 2001, vol. 2090/2001, pp. 1–37.

[53] J. Hillston, "Process algebras for quantitative analysis," in *Proc. IEEE International Symposium on Logic in Computer Science '05*, 2005, pp. 239–248.

[54] M. Molloy, "Performance analysis using stochastic Petri nets," *IEEE Trans. Comput.*, vol. C-31, no. 9, pp. 913–917, 1982.

[55] M. A. Marsan, G. Balbo, G. Conte, S. Donatelli, G. Franceschinis, and M. A. Marsan, *Modelling with Generalized Stochastic Petri Nets*. John Wiley & Sons, 1995.

[56] P. Haas and G. Shedler, "Regenerative stochastic Petri nets," *Elsevier Perform. Eval.*, vol. 6, no. 3, pp. 189–204, 1986.

[57] G. Balbo, "Introduction to Stochastic Petri Nets," in *Lectures on Formal Methods and Performance Analysis*, ser. Lecture Notes in Computer Science. Springer-Verlag, 2001, vol. 2090/2001, pp. 84–155.

[58] H. Hermanns, U. Herzog, and J.-P. Katoen, "Process algebra for performance evaluation," *Elsevier Theor. Comput. Sci.*, vol. 274, no. 1-2, pp. 43–87, 2002.

[59] A. Clark, S. Gilmore, J. Hillston, and M. Tribastone, "Stochastic Process Algebras," in *Formal Methods for Performance Evaluation*, ser. Lecture Notes in Computer Science. Springer-Verlag, 2007, vol. 4486/2007, pp. 132–179.

[60] J. C. M. Baeten, "A brief history of process algebra," *Elsevier Theoretical Computer Science*, vol. 335, no. 2-3, pp. 131–146, 2005.

[61] C. A. R. Hoare, *Communicating Sequential Processes*. Prentice Hall, 1985.

[62] R. Milner, *Communication and Concurrency*. Prentice Hall, 1989.

[63] J. Hillston, *A Compositional Approach to Performance Modelling*. Cambridge University Press, 1996.

[64] N. Gotz, U. Herzog, and M. Rettelbach, "TIPP—a language for timed processes and performance evaluation," University of Erlangen-Nurnberg, Tech. Rep., Nov. 1992.

[65] M. Bernardo and R. Gorrieri, "A tutorial on EMPA: a theory of concurrent processes with nondeterminism, priorities, probabilities and time," *Elsevier Theor. Comput. Sci.*, vol. 202, no. 1-2, pp. 1–54, 1998.

[66] G. D. Plotkin, "A Structural Approach to Operational Semantics," University of Aarhus, Tech. Rep., 1981.

[67] W. J. Stewart, *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, 1994.

[68] R. A. Howard, *Dynamic Probabilistic Systems: Volume 2, Semi-Markov and Decision Processes*. John Wiley & Sons, 1971.

[69] PEPA Tools. [Online]. Available: http://www.dcs.ed.ac.uk/pepa/tools/

[70] R. Huslende, "A combined evaluation of performance and reliability for degradable systems," in *Proc. ACM SIGMETRICS Conference on Measurements and Modeling of Computer Systems, Performance Evaluation Reviews '93*, 1981, pp. 157–164.

[71] H. Meer and H. Sevcikova, "PENELOPE dependability evaluation and the optimization of performability," in *Computer Performance Evaluation Modelling Techniques and Tools*, ser. Lecture Notes in Computer Science. Springer-Verlag, 1997, vol. 1245/1997, pp. 19–31.

[72] H. Meer, K. Trivedi, G. Bolch, and F. Hofmann, "Optimal transient service strategies for adaptive heterogeneous queueing systems," in *Proc. GI/ITG Conference on Measurement, Modelling and Performance Evaluation of Computer and Communication Systems '93*, 1993, pp. 166–170.

[73] N. M. van Dijk, *Queueing Networks and Product Forms: A Systems Approach*. John Wiley & Sons, 1993.

[74] M. Marsan, G. Ginella, R. R. Maglione, and M. Meo, "Performance analysis of hierarchical cellular networks with generally distributed call holding times and dwell times," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 248–257, 2004.

[75] H. Chaskar, "Requirements of a Quality of Service (QoS) Solution for Mobile IP," IETF RFC 3583, Sep. 2003.

[76] G.Chiruvolu, A. Agrawal, and M. Vandenhoute, "Mobility and QoS support for IPv6-based real-time wireless Internet traffic," in *Proc. IEEE International Conference on Communications '99*, 1999, pp. 334–338.

[77] G.-S. Kuo and P.-C. Ko, "Dynamic RSVP for mobile IPv6 in wireless networks," in *Proc. IEEE Vehicular Technology Conference-Spring '00*, 2000, pp. 455–459.

[78] X. Shangguan, W. Seah, and C. Ko, "Performance evaluation of a lightweight resource reservation protocol for mobile internet hosts," in *Proc. IEEE International Workshop on Mobile Computing Systems and Applications '00*, 2000, pp. 119–127.

[79] C. Shen, A. Lo, and W. Seah, "Performance evaluation of flow transparent mobile IPv6 and RSVP integration," in *Proc. Multi-Conference on Systemics, Cybernetics and Informatics '01*, 2001, pp. 515–520.

[80] S. Paskalis, A. Kaloxylos, E. Zervas, and L. Merakos, "Evaluating the RSVP mobility proxy concept," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications '02*, 2002, pp. 270–274.

[81] ——, "An efficient RSVP-mobile IP interworking scheme," *Springer Mob. Netw. Appl.*, vol. 8, no. 3, pp. 197–207, 2003.

[82] N.-F. Huang and W.-E. Chen, "RSVP extensions for real-time services in hierarchical mobile IPv6," *Springer Mob. Netw. Appl.*, vol. 8, no. 6, pp. 625–634, 2003.

[83] H.-W. Ferng, W.-Y. Kao, J.-J. Huang, and D. Shiung, "A dynamic resource reservation scheme designed for improving multicast protocols in HMIPv6-based networks," in *Proc. IEEE Vehicular Technology Conference-Spring '06*, 2006, pp. 961–965.

[84] A. Terzis, M. Srivastava, and L. Zhang, "A simple QoS signaling protocol for mobile hosts in the integrated services Internet," in *Proc. IEEE International Conference on Computer Communications '99*, 1999, pp. 1011–1018.

[85] G.-C. Lee, T.-P. Wang, and C.-C. Tseng, "Resource reservation with pointer forwarding schemes for the mobile RSVP," *IEEE Commun. Lett.*, vol. 5, no. 7, pp. 298–300, 2006.

[86] S.-C. Lo, G. Lee, W.-T. Chen, and J.-C. Liu, "Architecture for mobility and QoS support in all-IP wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 4, pp. 691–705, 2004.

[87] J. Rajahalme, A. Conta, B. Carpenter, and S. Deering, "IPv6 Flow Label Specification," IETF RFC 3697, Mar. 2004.

[88] J. Kempf, "Problem Statement for Network-Based Localized Mobility Management (NETLMM)," IETF RFC 4830, Apr. 2007.

[89] A. Terzis, J. Krawczyk, J. Wroclawski, and L. Zhang, "RSVP Operation Over IP Tunnels," IETF RFC 2746, Jan. 2000.

[90] D. Lam, D. Cox, and J. Widom, "Teletraffic modeling for personal communications services," *IEEE Commun. Mag.*, vol. 35, no. 2, pp. 79–87, 1997.

[91] T. Karagiannis, M. Molle, and M. Faloutsos, "Long-range dependence ten years of Internet traffic modeling," *IEEE Internet Comput.*, vol. 8, no. 5, pp. 57–64, 2004.

[92] D. Hong and S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Trans. Veh. Technol.*, vol. 35, no. 3, pp. 77–92, 1986.

[93] Y. Fang, I. Chlamtac, and Y.-B. Lin, "Modeling PCS networks under general call holding time and cell residence time distributions," *IEEE/ACM Trans. Netw.*, vol. 5, no. 6, pp. 893–906, 1997.

[94] Y.-B. Lin, "Modeling techniques for large-scale PCS networks," *IEEE Commun. Mag.*, vol. 35, no. 2, pp. 102–107, 1997.

[95] S. Kourtis and R. Tafazolli, "Evaluation of handover related statistics and the applicability of mobility modelling in their prediction," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications '00*, 2000, pp. 665–670.

[96] A. V. Aho, B. W. Kernighan, and P. J. Weinberger, *The AWK Programming Language*. Addison-Wesley Publishing Company, 1988.

[97] J. Hillston, "Fluid flow approximation of PEPA models," in *Proc. IEEE International Conference on Quantitative Evaluation of Systems '05*, 2005, pp. 33–43.

[98] S.-J. Leu and R.-S. Chang, "Integrated service mobile Internet: RSVP over mobile IPv4&6," *Springer Mob. Netw. Appl.*, vol. 8, no. 6, pp. 635–642, 2003.

[99] A. K. Talukdar, B. R. Badrinath, and A. Acharya, "MRSVP: a resource reservation protocol for an integrated services network with mobile hosts," *Springer Wirel. Netw.*, vol. 7, no. 1, pp. 5–19, 2001.

[100] C.-C. Tseng, G.-C. Lee, R.-S. Liu, and T.-P. Wang, "HMRSVP: a hierarchical mobile RSVP protocol," *Springer Wirel. Netw.*, vol. 9, no. 2, pp. 95–102, 2003.

[101] D. Awduche and E. Agu, "Mobile extensions to RSVP," in *Proc. IEEE International Conference on Computer Communications and Networks '97*, 1997, pp. 132–136.

[102] D. Levine, I. Akyildiz, and M. Naghshineh, "A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept," *IEEE/ACM Trans. Netw.*, vol. 5, no. 1, pp. 1–12, 1997.

[103] W.-T. Chen and L.-C. Huang, "RSVP mobility support: a signaling protocol for integrated services Internet with mobile hosts," in *Proc. IEEE International Conference on Computer Communications '00*, 2000, pp. 1283–1292.

[104] J. Manner, A. L. Toledo, A. Mihailovic, H. L. V. Munoz, E. Hepworth, and Y. Khouaja, "Evaluation of mobility and quality of service interaction," *Elsevier Comput. Netw.*, vol. 38, no. 2, pp. 137–163, 2002.

[105] G. Huston, "Next Steps for the IP QoS Architecture," IETF RFC 2990, Nov. 2000.

[106] S. Pack, X. S. Shen, J. W. Mark, and J. Pan, "Adaptive route optimization in hierarchical mobile IPv6 networks," *IEEE Trans. Mobile Comput.*, vol. 6, no. 8, pp. 903–914, 2007.

[107] S. Pack, T. Kwon, and Y. Choi, "A mobility-based load control scheme at mobility anchor point in hierarchical mobile IPv6 networks," in *Proc. IEEE Global Telecommunications Conference '04*, 2004, pp. 3431–3435.

[108] J. Markoulidakis, G. Lyberopoulos, D. Tsirkas, and E. Sykas, "Mobility modeling in third-generation mobile telecommunications systems," *IEEE Personal Commun. Mag.*, vol. 4, no. 4, pp. 41–56, 1997.

[109] C. Bettstetter, "Mobility modeling in wireless networks: categorization, smooth movement, and border effects," *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, no. 3, pp. 1559–1662, 2001.

[110] J. Almhana, Z. Liu, V. Choulakian, and R. McGorman, "A mobile terminal location tracking model for personal communication systems," in *Proc. IEEE International Conference on Local Computer Networks '05*, 2005, pp. 600–607.

[111] F. Baker, C. Iturralde, F. Faucheur, and B. Davie, "Aggregation of RSVP for IPv4 and IPv6 Reservations," IETF RFC 3175, Sep. 2001.

[112] M. Kibria and A. Jamalipour, "On designing issues of the next generation mobile network," *IEEE Netw.*, vol. 21, no. 1, pp. 6–13, 2007.

[113] J.-C. Cheng and T. Zhang, *IP-based Next-generation Wireless Networks*. John Wiley & Sons, 2004.

[114] "3GPP system to Wireless Local Area Network (WLAN) interworking; System description," 3GPP, Tech. Rep. TS23.234 v7.7.0, Jun. 2008.

[115] Q. Song and A. Jamalipour, "Network selection in an integrated wireless LAN and UMTS environment using mathematical modeling and computing techniques," *IEEE Wireless Commun. Mag.*, vol. 12, no. 3, pp. 42–48, 2005.

[116] W. Shen and Q.-A. Zeng, "Cost-function-based network selection strategy in integrated wireless and mobile networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 6, pp. 3778–3788, 2008.

[117] R. Skehill, M. Barry, W. Kent, M. O'Callaghan, N. Gawley, and S. Mcgrath, "The common RRM approach to admission control for converged heterogeneous wireless networks," *IEEE Wireless Commun. Mag.*, vol. 14, no. 2, pp. 48–56, 2007.

[118] A. Hasib and A. O. Fapojuwo, "Analysis of common radio resource management scheme for end-to-end QoS support in multiservice heterogeneous wireless networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 4, pp. 2426–2439, 2008.

[119] X. Gelabert, J. Perez-Romero, Q. Sallent, and R. Agusti, "A Markovian approach to radio access technology selection in heterogeneous multiaccess/multiservice wireless networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 10, pp. 1257–1270, 2008.

[120] J. Farber, S. Bodamer, and J. Charzinski, "Statistical evaluation and modelling of Internet dial-up traffic," in *Proc. SPIE Phonics East Conference: Performance and Control of Network Systems III*, 1999, pp. 112–121.

[121] H.-K. Choi and J. O. Limb, "A behavioral model of Web traffic," in *Proc. IEEE International Conference on Network Protocols '99*, 1999, pp. 327–334.

[122] Y. Sasaki, H. Imai, M. Tsunoyama, and I. Ishii, "Approximation of probability distribution functions by coxian distribution to evaluate multimedia systems," *Syst. Comput. Japan*, vol. 35, no. 2, pp. 16–24, 2004.

[123] A. H. Zahran, B. Liang, and A. Saleh, "Mobility modeling and performance evaluation of heterogeneous wireless networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 8, pp. 1041–1056, 2008.

[124] A. Thummler, P. Buchholz, and M. Telek, "A novel approach for phase-type fitting with the EM algorithm," *IEEE Trans. Dependable Secure Comput.*, vol. 3, no. 3, pp. 245–258, 2006.

[125] J. P. Romero, O. Sallent, R. Agusti, and M. A. Diaz-Guerra, *Radio Resource Management Strategies in UMTS*. John Wiley & Sons, 2005.

[126] J. Kalliokulju, P. Meche, M. J. Rinne, J. Vallstrom, P. Varshney, and S.-G. Haggman, "Radio access selection for multistandard terminals," *IEEE Commun. Mag.*, vol. 39, no. 10, pp. 116–124, 2001.

[127] Q.-A. Zeng and D. P. Agrawal, "Modeling and efficient handling of handoffs in integrated wireless mobile networks," *IEEE Trans. Veh. Technol.*, vol. 51, no. 6, pp. 1469–1478, 2002.

[128] H. Wang, D. Laurenson, and J. Hillston, "Process-algebra Based Framework using PEPA: Impact Assessment," Mobile VCE Core 4, Tech. Rep. U.2.3.4, Oct. 2007.

Appendix A
Publications

The author of this thesis has the following accepted or submitted publications during the course of his Ph.D. research:

A.1 Journal Papers

- Hao Wang, David I. Laurenson, and Jane Hillston, "A General Performance Evaluation Framework for Network Selection Strategies in 3G-WLAN Interworking Networks," submitted to *IEEE Transaction on Mobile Computing*. (on page 145)

- Hao Wang, David I. Laurenson, and Jane Hillston, "A Reservation Optimised Advance Resource Reservation Scheme for Deploying RSVP in Mobile Environments," *Springer Wireless Personal Communications*, published online first, 2009. (on page 161)

A.2 Conference Papers

- Hao Wang, David I. Laurenson, and Jane Hillston, "Evaluation of RSVP and Mobility-Aware RSVP Using Performance Evaluation Process Algebra," in *Proc. IEEE International Conference on Communications*, 2008, pp. 192–197. (on page 187)

- Hao Wang, David I. Laurenson, and Jane Hillston, "An SMR based advance resource reservation scheme for combined mobility and QoS Provisioning," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, 2008, pp. 1–5. (on page 192)

- Hao Wang, David I. Laurenson, and Jane Hillston, "PEPA Analysis of MAP Effects in Hierarchical Mobile IPv6," in *Proc. IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, 2007, pp. 337–342. (on page 197)

The original publications are included in the following pages.

A General Performance Evaluation Framework for Network Selection Strategies in 3G-WLAN Interworking Networks

Hao Wang, *Student Member, IEEE*, David I. Laurenson, *Member, IEEE*, and Jane Hillston, *Member, IEEE*

Abstract—In this work we investigate a general performance evaluation framework for network selection strategies (NSSs) that are used in 3G-WLAN interworking networks. Instead of simulation, this framework is based on models of NSSs and is constructed using a stochastic process algebra, named Performance Evaluation Process Algebra (PEPA). It captures the traffic and mobility characteristics of mobile nodes in 3G-WLAN interworking networks and has a good expression of the behaviour of the mobile nodes using different NSSs. Commonly used NSSs are evaluated from the perspectives of average throughput, handover rate, and network blocking probability. Results of the evaluation explore the effect of these NSSs on both mobile nodes and networks, as well as their characteristics in different mobility and traffic scenarios.

Index Terms—3G, WLAN, network selection, performance evaluation, Markov chain, process algebra.

1 Introduction

With the rapid development of various wireless communication technologies, the main trend for future wireless communications is a shift from voice and text based services provided by early cellular networks to multimedia-based services provided by multiple and heterogeneous wireless networks. These multimedia applications require different bearer services and there is no single technology that can simultaneously provide low latency, cheap cost, high data rate and high mobility to mobile users [1]. Therefore, next-generation wireless communications focuses on the integration of existing cellular and other wireless networks. Especially, owing to the recent evolution and successful deployment of the wireless local area network (WLAN), there has been a demand for integrating WLAN with the third-generation (3G) cellular network, i.e. 3G-WLAN interworking. For example, the 3rd generation partnership project (3GPP), an international organisation which produces technical specifications and reports for 3G wireless wide area networks (WWANs), has developed an architecture for the integration of the 3GPP cellular network and WLAN with the intention to extend 3GPP services to the WLAN access environment [2].

In a 3G-WLAN interworking network, a mobile node may perform vertical handovers (VHOs) [3] when it moves across overlaid 3G and WLAN radio access networks (RANs). During the handover, the mobile node may use its network selection strategy (NSS) [4] to select a certain type of radio access technology (RAT) to continue its communication. Various handover criteria can be taken into account when making a vertical handover decision, e.g., cost of services, network and mobile node conditions, and user preference, etc [5]. NSSs have been designed to abstract the process of selecting an appropriate network as mathematical problems which assess alternative networks in terms of different criteria. For a general review of the proposed strategies and their details, refer to [6], [7] and the references therein. Since NSSs control the session behaviour of mobile nodes, it is important and meaningful to investigate how they affect the performance of user applications at the mobile nodes and the utilisation of the RANs. Moreover, it would be better if the assessment can be conducted using a formal performance modelling technique rather than system-level simulations.

In this work, we investigate a general performance evaluation framework for NSSs in 3G-WLAN interworking networks. The framework is based on the performance models constructed using a formal performance modelling technique, named Performance Evaluation Process Algebra (PEPA). PEPA is chosen because its behavioural and compositional system description capability can efficiently reduce the modelling and evaluation difficulty of complex systems, and therefore facilitate the whole evaluation process. This framework captures the traffic and mobility characteristics of mobile nodes in 3G-WLAN interworking networks and has a good expression of the behaviour of the mobile nodes using different NSSs. From the framework, important performance measures including average throughput, RAN blocking probability and handover rate are derived.

The rest of this paper is structured as follows. Section 2 gives a brief review of the previous work on evaluating NSSs used in 3G-WLAN interworking networks. A traffic model of multimedia communications and a

This research has been funded by the Industrial Companies who are Members of Mobile VCE, with additional financial support from the UK Governments Technology Strategy Board (previously DTI).

mobility model suitable for 3G-WLAN interworking networks which are used in the framework are presented in Section 3 and Section 4 respectively. In Section 5, a short introduction to the PEPA formalism is given and in Section 6 PEPA models of different NSSs are described. In Section 7, an iterative method that interlinks the PEPA models and a network resource consumption model in order to derive the RAN blocking probability and handover rate is described. The performance of different NSSs are evaluated in Section 8 and in Section 9 the evaluation results are discussed.

2 Related Work and Contributions

In previous studies on the design and evaluation of NSSs in 3G-WLAN interworking networks, the selection strategies are first formulated as policy functions which take network and mobile node conditions as the input parameters and then output the preferred RAN. Then comparisons are made between the results of different strategies, or between the results of the same strategy under different scenarios, in terms of throughput, connection cost, power consumption, etc. Typical examples of this can be found in [6], [8], [9]. There are also studies which aim to integrate resource management schemes into the NSSs. In [10], network selection takes account of the traffic loads of different RANs in order to achieve a balance between the RANs. In [11], [12], the mobility, traffic and location information of the mobile nodes are considered, and network selection is implemented in a centralised way to manage the resources of the RANs more efficiently. An analytical model for evaluating different NSSs is investigated in [13]. Each strategy is formulated as a mapping function between the total session arrival rates of the whole 3G-WLAN interworking network and the session arrival rates into different RANs. However, this work only considers network selection for new session requests and node mobility is not taken into account.

One major limitation of the above studies is that the evaluation is usually carried out in a restricted way. That is, different strategies are compared at the policy function level without considering the mobile node's session and handover behaviour, whereas evaluation with consideration of traffic and mobility only focuses on a certain type of strategy. However, to make a fair comparison, different strategies should be evaluated using the same general framework with as few restrictions as possible. To the best of our knowledge, there are no such framework for evaluating NSSs in the published literature, and this work is thus motivated by investigating such a general performance evaluation framework. The contributions of this work are summarised as follows:

- With the help of the behavioural and compositional system description capability of the PEPA language, in the framework, different NSSs are modelled in terms of the behaviour of a mobile node. To make the description more accurate, a traffic model which represents features of multimedia-based services, and a mobility model which captures movement characteristics of a mobile node, are employed.
- NSSs for both new and handover sessions are considered in the framework, and they are embedded in the form of network selection probabilities. This level of abstraction provides the framework the flexibility to express any type of NSS given that its probabilities of choosing RANs can be determined.
- The generality of the framework is also retained by having an interface to the model capturing RAN resource consumption, in the form of the network blocking probability. In this way, the framework is independent on the call admission control (CAC) and resource management schemes used in the 3G-WLAN interworking networks.
- Moreover, a novel iterative algorithm is proposed to derive network blocking probability from this framework when it is not at the modellers' disposal. This algorithm links models of RAN resource consumption and models of different NSSs through interchanging necessary parameters between them. The convergence speed of the algorithm is fast, and performance measures are direct results of it.

3 Traffic Model

The traffic of a mobile node is modelled at the session level in the framework, which includes two parameters: session arrival rate and session duration. Due to the multi-service characteristic of next-generation communications, the mobile node requires different types of services that have different durations. This combination of various services results in a large variance in the observed user session duration. Moreover, the field data suggests that the statistical session duration in the Internet has a coefficient of variation (CoV) larger than one [14], [15]. To capture this traffic feature, we use the hyper-exponential distribution to model the session duration of a mobile node.

A K-phase ($K > 1$) hyper-exponential distribution is composed of K exponentially distributed phases in parallel and always has its CoV larger than one. To simplify the traffic model, a two-phase hyper-exponential distribution is used, where one phase represents the non-real time (NRT) sessions and the other represents the real-time (RT) sessions. The NRT sessions generally include Web surfing and file transfer, and the RT sessions generally include voice and video streaming. As for the session arrival rate, the general consensus that the session arrival is a Poisson process is followed.

On the basis of the above assumptions, the traffic model can be constructed as a combination of two ON-OFF sources. As shown in Fig 1, a session arrives at the rate of λ and it can be either an RT session with probability P_{RT} or an NRT session with probability $P_{NRT} = 1 - P_{RT}$. The mean durations of the RT and NRT sessions are $1/\mu_{RT}$ and $1/\mu_{NRT}$ respectively.

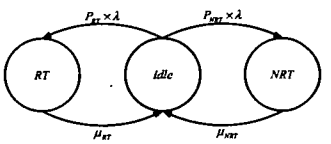


Fig. 1. A traffic model with two ON-OFF sources

4 Mobility Model

In a 3G-WLAN interworking network, a cellular cell is usually called a *3G-WLAN interworking cell*, and is generally overlaid with one or more WLAN cells. Therefore, the mobility model of a mobile node should characterise its residence times not only in the whole 3G-WLAN interworking cell but also in the different RAT areas within the 3G-WLAN interworking cell. To this end, the mobility of the mobile node in the 3G-WLAN interworking network is modelled as a continuous-time Markov chain (CTMC) which has the *Coxian* structure.

A K-phase Coxian structure is composed of a series of K exponentially distributed phases (or states) and an absorbing phase. Phase i either enters the phase $i+1$ or enters the absorbing phase with pre-defined probabilities. Phase K enters the absorbing phase with the probability of 1. A Coxian distribution is then defined as the distribution from the first phase until absorption, and it has an important property that it can arbitrarily closely approximate any probability distribution [16]. By letting the phases represent the position of the mobile node in a 3G-WLAN interworking cell in terms of the RAT area, the transitions of the Coxian structure can capture various tracks that the mobile node may traverse in a 3G-WLAN interworking cell [17].

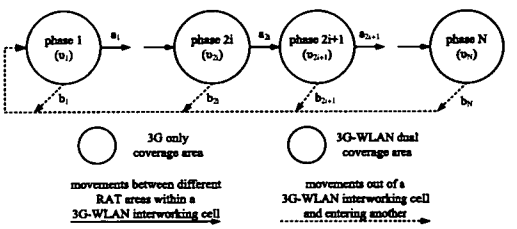


Fig. 2. A mobility model with the Coxian structure

Since the ordinary Coxian structure contains an absorbing state whereas the focus of this work is on the steady state performance of a mobile node, the absorbing state is omitted in the mobility model. The mobility model with the Coxian structure is assumed to have an even number (N) of phases, and it is shown in Fig. 2. The odd phases $2i-1$ and the even phases $2i$ represent the mobile node being in the *3G only coverage areas* and in the *3G-WLAN dual coverage areas* respectively, where $i = 1, 2, \ldots, N/2$. Without ambiguity, the above two types of area are called a *3G area* and a *3G-WLAN area* respectively. Two important assumptions are made in this work: Firstly, WLAN cells are assumed not to overlap with each other. Secondly, a WLAN cell which overlaps with two adjacent cellular cells is considered to belong to *both* cellular cells. With the second assumption, the starting point of the track of the mobile node in a 3G-WLAN interworking cell is always the 3G area.

The state transition diagram shown in Figure 2 captures the movement of a mobile node within and across the 3G-WLAN interworking cells: Movements across different RAT areas *within* a 3G-WLAN interworking cell are captured by the transitions between neighbouring phases. Movements *out of* the 3G-WLAN interworking cell are captured by the transitions from interim phases back to the first phase, and mean the mobile node enters the 3G area of *another new* 3G-WLAN interworking cell.

Owing to the Coxian structure of the mobility model, the sojourn time of the mobile node in the 3G-WLAN interworking cell, i.e., the time spent by the mobile node traversing a series of phases until going back to the first phase, still follows the Coxian distribution. In the mobility model, each phase of the mobility model, e.g. phase k, is assumed to be exponentially distributed with rate v_k, and the probabilities of branching to the next and the first phases are a_k and b_k respectively. The number of phases, the rate and the branching probabilities of each phase can be estimated from field data using algorithms presented in [16], [18].

5 Performance Evaluation Process Algebra

In this section, we give a short introduction to the PEPA language which is adopted as the modelling technique in this work.

Process algebras have been designed as formal description techniques for concurrent systems — systems that consist of subsystems interacting with each other. In process algebras, a *process* or an *agent* can perform *actions*, and a system is modelled from the perspective of its *behaviour* as interactions between processes. PEPA [19] is a stochastic extension of classical process algebras. It is important to emphasise that PEPA is not a Monte-Carlo evaluation approach but an advanced mathematical modelling technique that is ideal for performance modelling of large and complex resource-sharing systems.

5.1 Syntax of PEPA

PEPA is a compositional approach that decomposes a system into subsystems that are smaller and more easily modelled. In PEPA a system is usually modelled to be composed of a group of *components* that engage in *activities*. This abstract description of a system is similar to the design process of a system and facilitates model construction. Generally, components model the physical or logical elements of a system and activities characterise

the behaviour of these components. For example, a printing system can be considered to consist of a *Queue* component which buffers jobs and a *Printer* component which prints jobs.

Each activity a in PEPA is defined as a pair (α, r) — the action type α, which can be regarded as the name of the activity, and the activity rate r, which is the parameter of an exponentially distributed random variable that specifies the duration of the activity. If a component P behaves as P' after completing activity a, then we can regard this behaviour as a component changing from state P to state P', through transition (α, r).

The PEPA formalism provides a small set of operators which are able to express the individual activities of components as well as the interactions between them. Here we only present the operators that are used in this work. For more details about PEPA operators, see [19].

Prefix: $(\alpha, r).P$

This component has a designated first activity which is of action type (or name) α and has a duration that is exponentially distributed with rate r, which gives a mean time of $1/r$. After completing this activity, the component $(\alpha, r).P$ behaves as P. For example, the *Printer* component in the above example can print a job and then suspend for a while before it is ready to take the next job. This behaviour can be expressed as:

$$Printer \stackrel{def}{=} (print, r_1).(suspend, r_2).Printer$$

Choice: $P + Q$

This component may either behave as P or Q. All the enabled activities in P and Q are enabled in this component and compete with each other. The first activity to be completed will be an activity of P or Q and this will distinguish which component wins the race. When the first activity is completed, all the other activities will be abandoned. For example, let the component $Queue_i$ denote the behaviour of the *Queue* component when there are i jobs in the queue. It can either allow another job to arrive (when the queue is not full) or have one of its jobs printed (when the queue is not empty). The *Queue* component can then be defined as $(i = 1, 2, \cdots, N-1)$:

$$\begin{aligned} Queue_0 &\stackrel{def}{=} (arrive, r_3).Queue_1 \\ Queue_i &\stackrel{def}{=} (arrive, r_3).Queue_{i+1} \\ &\quad + (print, \top).Queue_{i-1} \\ Queue_N &\stackrel{def}{=} (print, \top).Queue_{N-1} \end{aligned}$$

where N is the maximum size of the queue. The symbol "$\top$" means the rate of the activity is outside control of the component. In this example, the *Queue* component is *passive* with respect to the activity *print* since it cannot influence the rate at which jobs are printed.

Cooperation: $P \underset{L}{\bowtie} Q$

This component represents the interaction between P and Q. The set L is called the *cooperation set* and denotes a set of action types that must be carried out by P and Q together. For all activities whose action type is included in set L, P and Q must cooperate to complete it. However, other activities of P and Q which have types that are not included in set L will proceed independently. The rate of the *shared* activity is determined by the rate of the slower participant and is the smaller of the two rates. When the cooperation set L is empty, the two components proceed concurrently without any interaction between them. A shorthand notation $P \| Q$ is used to represent $P \underset{\varnothing}{\bowtie} Q$. For example, the printing system may have two parallel queues that share only one printer, and each *Queue* component cooperates with the *Printer* component on the activity *print* individually. This can be expressed as:

$$(Queue_0 \| Queue_0) \underset{\{print\}}{\bowtie} Printer$$

Constant: $P \stackrel{def}{=} Q$

The constant operator "$\stackrel{def}{=}$" can be used to associate names with behaviour. Its usage to define single components has been shown in the above examples. Moreover, it can be used for *system definition* which specifies how the system is constructed from the defined components, i.e., how the components cooperate with each other so that they express the behaviour of the system. For example, the printing system with two queues and one printer can be given a name *System*, which is associated with the cooperation between the components *Queue* and *Printer*. That is:

$$System \stackrel{def}{=} (Queue_0 \| Queue_0) \underset{\{print\}}{\bowtie} Printer$$

where $Queue_0$ and *Printer* define the initial behaviour of the corresponding components.

The system states of a PEPA model are the *feasible* combinations of the state of each component of that model. The above printing system example with two queues and one printer has 32 system states if N is 3. For example, the state $(Queue_2 \| Queue_1) \underset{\{print\}}{\bowtie} Printer'$ is one of them.

5.2 Deriving Performance Measures

For any PEPA model, an underlying stochastic process can be generated: a state is associated with a component, and the transitions between the states are defined by the activities between them. Since the duration of a transition in PEPA is exponentially distributed, it has been established that the stochastic process underlying a PEPA model is a CTMC. Performance measures are usually derived from the equilibrium probability vector of the CTMC using the Markov reward structure (MRS) [20].

The equilibrium probabilities of a CTMC can be regarded as the probability that the system is in a state when observed at a random time, or alternatively, can be regarded as the portion of the time the system spends in that state. In MRSs, rewards are associated with states of a Markov process or with transitions between

states, and performance measures are then calculated as the total reward based on the equilibrium system state distribution. If ρ_i is the reward associated with system state s_i, and $\pi(s_i)$ is the equilibrium probability of s_i, then the total reward R is:

$$R = \sum_{s_i \in S} \rho_i * \pi(s_i), \quad (1)$$

where S is the set of all the feasible system states of the PEPA model and ρ_i can be any meaningful value. In this way, various types of measures can be derived from the PEPA model level rather than at the Markov process level.

6 PEPA Models of Network Selection Strategies

In this section, PEPA models that describe the behaviour of a mobile node using different NSSs are presented. Each strategy has its own corresponding PEPA model, and each model consists of three PEPA components that capture the session, handover and network selection behaviour of the mobile node. The same traffic and mobility patterns of the mobile node are used in each model in order that the results depend only on the characteristics of different NSSs. In the following subsections, the PEPA components for the session and handover behaviour of the mobile node are presented first, followed by the PEPA components corresponding to each NSS.

6.1 PEPA Component for Traffic Model

A mobile node's session behaviour is modelled by the component *SESS*. According to Section 3, it can be in an idle or an engaged state.

6.1.1 Idle State of Component SESS

In state $SESS_{Idle}$, the component *SESS* can carry out two sets of new session request activities, depending on the position of the mobile node in the 3G-WLAN interworking cell:

- When the mobile node is in the 3G area, new session requests are generated at the rate of λ and all of them are submitted to the 3G RAN (3GRAN). The new session request activities are classified by the session type (activities $sess\_req_{NRT}$ and $sess\_req_{RT}$) and they are generated with the probabilities P_{NRT} and P_{RT} respectively.
- When the mobile node is in the 3G-WLAN area, the new session request activities are further classified by the RAN the requests are submitted to. For example, the activities $sess\_req^{C}_{NRT}$ and $sess\_req^{W}_{RT}$ mean that an NRT session request is submitted to the 3GRAN and an RT session request is submitted to the WLAN RAN (WRAN) respectively. The RAN to which the request is submitted is determined by the PEPA component for the NSS, which will be discussed in Section 6.3.
- Once the new session requests are admitted, the component *SESS* goes to one of the engaged states. The probability of admitting the new session requests is reflected in the cooperation between the component *SESS* and the PEPA component for mobility, which will be discussed in Section 6.2.[1] A new session request may be rejected, and in this case the component *SESS* remains in the idle state. Since the activities corresponding to new session blocking are self-transitions and have no effect on the equilibrium probability of the system states, they are omitted and not defined in the idle state.

The idle state of the component *SESS* is defined as:

$$\begin{aligned} SESS_{Idle} \stackrel{def}{=} & (sess\_req_{NRT},\ P_{NRT} * \lambda).SESS_{NRT} \\ + & (sess\_req_{RT},\ P_{RT} * \lambda).SESS_{RT} \\ + & (sess\_req^{C}_{NRT},\ P_{NRT} * \lambda).SESS_{NRT} \\ + & (sess\_req^{W}_{NRT},\ P_{NRT} * \lambda).SESS_{NRT} \\ + & (sess\_req^{C}_{RT},\ P_{RT} * \lambda).SESS_{RT} \\ + & (sess\_req^{W}_{RT},\ P_{RT} * \lambda).SESS_{RT} \end{aligned}$$

6.1.2 Engaged States of Component SESS

Depending on the type of a session, the component *SESS* can be either in the state $SESS_{NRT}$ or in the state $SESS_{RT}$. When the session is completed (activities $sess_{NRT}$ and $sess_{RT}$), or is dropped during a handover (activities *HHO_blk* and *VHO_blk*), the component *SESS* goes back to the idle state.

The engaged states of the component *SESS* are defined as:

$$\begin{aligned} SESS_{NRT} \stackrel{def}{=} & (sess_{NRT},\ \mu_{NRT}).SESS_{Idle} \\ + & (VHO\_blk,\ \top).SESS_{Idle} \\ + & (HHO\_blk,\ \top).SESS_{Idle} \\ SESS_{RT} \stackrel{def}{=} & (sess_{RT},\ \mu_{RT}).SESS_{Idle} \\ + & (VHO\_blk,\ \top).SESS_{Idle} \\ + & (HHO\_blk,\ \top).SESS_{Idle} \end{aligned}$$

6.2 PEPA Component for Mobility Model

A mobile node's handover behaviour is captured by the component *MN*. Moreover, the component *MN* also expresses the session and network selection behaviour of the mobile node, but they are *controlled* by the corresponding PEPA components, as reflected by the symbol "$\top$". Like the component *SESS*, the component *MN* can be in an idle or an engaged state.

6.2.1 Idle States of Component MN

The idle states of the component *MN* imply that the mobile node is not communicating and they are characterised by the position of the mobile node in the 3G-WLAN interworking cell. The subscripts $2i-1$ and $2i$ ($i = 1, 2, \ldots, N/2$) are used to denote which phase of the mobility model the mobile node is currently in. The idle states of the component *MN* cooperate with the

1. The activity rates of successful new session requests are in fact not $P_{NRT} * \lambda$ and $P_{RT} * \lambda$ and they are determined by the cooperation.

component *SESS* on different sets of new session request activities according to the mobile node's position:

- In state MN^{Idle}_{2i-1}, the mobile node is in the 3G area. It submits new session requests to the 3GRAN (activities $sess\_req_{NRT}$ and $sess\_req_{RT}$). The probability of admitting a new session request in the 3GRAN is reflected in the parameter P^C_{NA}. For example, by the cooperation between MN^{Idle}_{2i-1} and $SESS_{Idle}$, the rate of the activity $sess\_req_{NRT}$ is actually $P^C_{NA} * P_{NRT} * \lambda$, rather than $P_{NRT} * \lambda$.
- In state MN^{Idle}_{2i}, the mobile node is in the 3G-WLAN area. It submits new session requests to different RANs according to its selection strategy (e.g. activities $sess\_req^C_{NRT}$ and $sess\_req^W_{RT}$). The probability of admitting a new session request to the WRAN is P^W_{NA}. The probabilities of selecting 3GRAN and WRAN are P_C and P_W respectively and they are defined in the PEPA component for NSS. For example, the rate of the activity $sess\_req^W_{RT}$ is actually $P^W_{NA} * P_W * P_{RT} * \lambda$.
- Once new session requests are admitted, the component *MN* goes into an engaged state. As for the component *SESS*, the activities corresponding to new session blocking are omitted since they are self-transitions.
- The mobile node can stay in the idle state and just move within and across the 3G-WLAN interworking cells. The activity $move_{m,n}$ represents the movement of the mobile node from phase m to phase n of its mobility model.

The idle states of the component *MN* are defined as ($i = 1, 2, \ldots, N/2$):

$$
\begin{aligned}
MN^{Idle}_{2i-1} \stackrel{def}{=}\ & (sess\_req_{NRT},\ P^C_{NA} * \top).MN^C_{2i-1} \\
+\ & (sess\_req_{RT},\ P^C_{NA} * \top).MN^C_{2i-1} \\
+\ & (move_{2i-1,2i},\ a_{2i-1} * \upsilon_{2i-1}).MN^{Idle}_{2i} \\
+\ & (move_{2i-1,1},\ b_{2i-1} * \upsilon_{2i-1}).MN^{Idle}_{1} \\
MN^{Idle}_{2i} \stackrel{def}{=}\ & (sess\_req^C_{NRT},\ P^C_{NA} * \top).MN^C_{2i} \\
+\ & (sess\_req^C_{RT},\ P^C_{NA} * \top).MN^C_{2i} \\
+\ & (sess\_req^W_{NRT},\ P^W_{NA} * \top).MN^W_{2i} \\
+\ & (sess\_req^W_{RT},\ P^W_{NA} * \top).MN^W_{2i} \\
+\ & (move_{2i,2i+1},\ a_{2i} * \upsilon_{2i}).MN^{Idle}_{2i+1} \\
+\ & (move_{2i,1},\ b_{2i} * \upsilon_{2i}).MN^{Idle}_{1} \\
MN^{Idle}_{N} \stackrel{def}{=}\ & (sess\_req^C_{NRT},\ P^C_{NA} * \top).MN^C_{N} \\
+\ & (sess\_req^C_{RT},\ P^C_{NA} * \top).MN^C_{N} \\
+\ & (sess\_req^W_{NRT},\ P^W_{NA} * \top).MN^W_{N} \\
+\ & (sess\_req^W_{RT},\ P^W_{NA} * \top).MN^W_{N} \\
+\ & (move_{N,1},\ b_N * \upsilon_N).MN^{Idle}_{1}
\end{aligned}
$$

6.2.2 Engaged States of Component MN

The engaged states of the component *MN* imply that the mobile node is communicating and they are characterised by both the position of the mobile node in the 3G-WLAN interworking cell and the RAN it is currently connected to. The superscripts C and W are used to denote the mobile node is using the 3GRAN and WRAN respectively. The engaged states of the component *MN* cooperate with the component *SESS* on the session holding activities and with the component for the NSS on the network selection activities.

- In state MN^C_{2i-1}, the mobile node is in the 3G area and is connected to the 3GRAN. When it moves into another 3G-WLAN interworking cell (i.e. from phase $2i-1$ to phase 1), it performs a horizontal handover (HHO) (activity $HHO_{2i-1,1}$). When it moves into the 3G-WLAN area of the same cell (i.e. from phase $2i-1$ to phase $2i$), it either performs a VHO to the WRAN (activity $VHO_{2i-1,2i}$), or performs no handover (NHO) and keeps its connection to the 3GRAN (activity $NHO_{2i-1,2i}$). Which type of handover is performed is decided by its NSS and is controlled by the PEPA component for NSS.[2]
- In state MN^C_{2i}, the mobile node is in the 3G-WLAN area and is connected to the 3GRAN. When it moves into another 3G-WLAN interworking cell (i.e. from phase $2i$ to phase 1), it performs a HHO (activity $HHO_{2i,1}$). When it moves into the 3G area of the same 3G-WLAN interworking cell (i.e. from phase $2i$ to phase $2i+1$), no handover is required (activity $NHO_{2i,2i+1}$).
- In state MN^W_{2i}, the mobile node is in the 3G-WLAN area and is connected to the WRAN. It always performs a VHO (activities $VHO_{2i,1}$ and $VHO_{2i,2i+1}$) when it moves out of the 3G-WLAN area (i.e. from phase $2i-1$ to phase 1 and from phase $2i-1$ to phase $2i$). To help understand the above different types of handovers, Fig. 3 illustrates the handover related transitions between the engaged states.

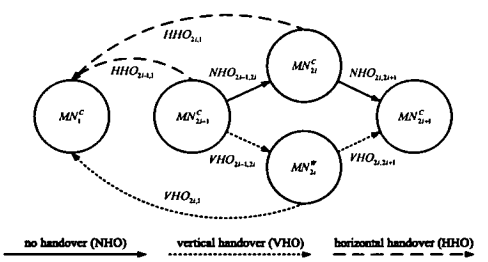


Fig. 3. Different types of handovers between the engaged states

- When the mobile node finishes its session (activities $sess_{NRT}$ and $sess_{RT}$), the component *MN* returns to an idle state. Generally, the NRT sessions are aware of the different data rates provided by different RATs. Therefore, the factors R^C_{NRT} and R^W_{NRT} are used to adjust the duration of the NRT sessions when the mobile node is connected to the 3GRAN and WRAN respectively. For example, the duration of the NRT session in the 3GRAN is $1/(R^C_{NRT} * \mu_{NRT})$.

2. Note that since the network selection only happens when the mobile node moves from the 3G area into the 3G-WLAN area, only $NHO_{2i-1,2i}$, $VHO_{2i-1,2i}$ and *VHO_blk* are *network selection related* activities during a handover.

- When the horizontal and vertical handover requests of the mobile node are blocked (activities *HHO_blk* and *VHO_blk*), the component *MN* also goes back to an idle state. The probabilities of admitting and blocking handover sessions in the 3GRAN and WRAN are P_{HA}^{C}, P_{HB}^{C}, P_{HA}^{W}, and P_{HB}^{W} respectively.

The engaged states of the component *MN* are defined as ($i = 1, 2, \ldots, N/2$):

$$
\begin{aligned}
MN_{2i-1}^{C} \stackrel{def}{=}\ & (sess_{NRT},\ R_{NRT}^{C} * \top).MN_{2i-1}^{Idle} \\
+\ & (sess_{RT},\ \top).MN_{2i-1}^{Idle} \\
+\ & (HHO_{2i-1,1},\ P_{HA}^{C} * b_{2i-1} * \upsilon_{2i-1}).MN_{1}^{C} \\
+\ & (HHO\_blk,\ P_{HB}^{C} * b_{2i-1} * \upsilon_{2i-1}).MN_{1}^{Idle} \\
+\ & (NHO_{2i-1,2i},\ a_{2i-1} * \upsilon_{2i-1}).MN_{2i}^{C} \\
+\ & (VHO_{2i-1,2i},\ P_{HA}^{W} * a_{2i-1} * \upsilon_{2i-1}).MN_{2i}^{W} \\
+\ & (VHO\_blk,\ P_{HB}^{W} * a_{2i-1} * \upsilon_{2i-1}).MN_{2i}^{Idle} \\
MN_{2i}^{C} \stackrel{def}{=}\ & (sess_{NRT},\ R_{NRT}^{C} * \top).MN_{2i}^{Idle} \\
+\ & (sess_{RT},\ \top).MN_{2i}^{Idle} \\
+\ & (HHO_{2i,1},\ P_{HA}^{C} * b_{2i} * \upsilon_{2i}).MN_{1}^{C} \\
+\ & (HHO\_blk,\ P_{HB}^{C} * b_{2i} * \upsilon_{2i}).MN_{1}^{Idle} \\
+\ & (NHO_{2i,2i+1},\ a_{2i} * \upsilon_{2i}).MN_{2i+1}^{C} \\
MN_{2i}^{W} \stackrel{def}{=}\ & (sess_{NRT},\ R_{NRT}^{W} * \top).MN_{2i}^{Idle} \\
+\ & (sess_{RT},\ \top).MN_{2i}^{Idle} \\
+\ & (VHO_{2i,1},\ P_{HA}^{C} * b_{2i} * \upsilon_{2i}).MN_{1}^{C} \\
+\ & (VHO\_blk,\ P_{HB}^{C} * b_{2i} * \upsilon_{2i}).MN_{1}^{Idle} \\
+\ & (VHO_{2i,2i+1},\ P_{HA}^{C} * a_{2i} * \upsilon_{2i}).MN_{2i+1}^{C} \\
+\ & (VHO\_blk,\ P_{HB}^{C} * a_{2i} * \upsilon_{2i}).MN_{2i+1}^{Idle} \\
MN_{N}^{C} \stackrel{def}{=}\ & (sess_{NRT},\ R_{NRT}^{C} * \top).MN_{N}^{Idle} \\
+\ & (sess_{RT},\ \top).MN_{N}^{Idle} \\
+\ & (HHO_{N,1},\ P_{HA}^{C} * \upsilon_{N}).MN_{1}^{C} \\
+\ & (HHO\_blk,\ P_{HB}^{C} * \upsilon_{N}).MN_{1}^{Idle} \\
MN_{N}^{W} \stackrel{def}{=}\ & (sess_{NRT},\ R_{NRT}^{W} * \top).MN_{N}^{Idle} \\
+\ & (sess_{RT},\ \top).MN_{N}^{Idle} \\
+\ & (VHO_{N,1},\ P_{HA}^{C} * \upsilon_{N}).MN_{1}^{C} \\
+\ & (VHO\_blk,\ P_{HB}^{C} * \upsilon_{N}).MN_{1}^{Idle}
\end{aligned}
$$

6.3 PEPA Component for Each Network Selection Strategy

A mobile node's network selection behaviour is controlled by the component *NS*, and a different component *NS* is required for each strategy. The component *NS* is designed to synchronise with the component *SESS* and the component *MN* on the *network selection related* activities. In this way, the mobile node's choice of a certain RAN can easily be controlled by enabling and disabling the synchronisation of the corresponding activities.

Different NSSs are classified into two groups: non-deterministic and deterministic. Non-deterministic strategies choose the RAN according to some on-line measures. On the other hand, deterministic strategies choose the RAN according to a predefined procedure. In this work, PEPA models of three types of selection strategies are built, namely general, WLAN-first and service-based. The general NSS model is for the common non-deterministic strategies and it embeds randomness in network selection. The WLAN-first and service-based NSSs models are for the specific deterministic strategies as their names suggest. The component *NS* and the system definition for each type of strategy are described in the following subsections.

6.3.1 General Network Selection Strategy

Using the general strategy, the mobile node chooses the 3GRAN and WRAN with non-zero probabilities P_C and P_W respectively. One example is the random strategy which chooses the 3GRAN and the WRAN with equal probabilities. Another example is the strategy that is based on the relative received signal strength (RRSS) [10]. No matter how the non-deterministic strategies are designed, the PEPA component for the general network selection NS^G should enable selecting both RANs for new and handover sessions. The component NS^G is defined as:

$$
\begin{aligned}
NS^{G} \stackrel{def}{=}\ & (sess\_req_{NRT}^{C},\ P_C * \top).NS^{G} \\
+\ & (sess\_req_{RT}^{C},\ P_C * \top).NS^{G} \\
+\ & (sess\_req_{NRT}^{W},\ P_W * \top).NS^{G} \\
+\ & (sess\_req_{RT}^{W},\ P_W * \top).NS^{G} \\
+\ & (NHO_{2i-1,2i},\ P_C * \top).NS^{G} \\
+\ & (VHO_{2i-1,2i},\ P_W * \top).NS^{G} \\
+\ & (VHO\_blk,\ P_W * \top).NS^{G}
\end{aligned}
$$

6.3.2 WLAN-first Network Selection Strategy

Using the WLAN-first strategy, the mobile node always chooses the WRAN whenever it is available. WRAN is usually preferred because of its high bandwidth, small delay and low cost. This strategy can be implemented by disabling the activities corresponding to selecting the 3GRAN for both new and handover sessions. The PEPA component for WLAN-first network selection NS^{WF} is defined as:

$$
\begin{aligned}
NS^{WF} \stackrel{def}{=}\ & (sess\_req_{NRT}^{W},\ P_W * \top).NS^{WF} \\
+\ & (sess\_req_{RT}^{W},\ P_W * \top).NS^{WF} \\
+\ & (VHO_{2i-1,2i},\ P_W * \top).NS^{WF} \\
+\ & (VHO\_blk,\ P_W * \top).NS^{WF}
\end{aligned}
$$

6.3.3 Service-based Network Selection Strategy

Using the service-based strategy, the mobile node chooses the RAN according to the type of its ongoing session. For example, the mobile node may choose the WRAN for NRT sessions and 3GRAN for RT sessions, because the NRT sessions can take advantage of the higher data rate provided by WRAN and the RT sessions will experience less handovers when choosing 3GRAN.

Since the mobile node makes its decision based on the session type, it would be better to let the component *SESS* control the network selection behaviour. For that reason, the engaged states of the component *SESS* are modified to implement service-based network selection during *handover*. For the NRT sessions, since the mobile node always performs a VHO from the 3GRAN to the WRAN, the activity $VHO_{2i-1,2i}$ is enabled. On the other hand, since the RT sessions always use the 3GRAN and no vertical handover is required, the activity $NHO_{2i-1,2i}$ is enabled and the activity *VHO_blk* is not needed. The component *SESS* for the service-based strategy is modified and renamed as $SESS^{SB}$, whose idle state is

the same as that of *SESS* whereas its engaged states are defined as:

$$\begin{aligned} SESS^{SB}_{NRT} &\stackrel{def}{=} (sess_{NRT},\ \mu_{NRT}).SESS^{SB}_{Idle} \\ &+ (VHO_{2i-1,2i},\ P_W * \top).SESS^{SB}_{NRT} \\ &+ (VHO\_blk,\ P_W * \top).SESS^{SB}_{Idle} \\ &+ (HHO\_blk,\ \top).SESS^{SB}_{Idle} \\ SESS^{SB}_{RT} &\stackrel{def}{=} (sess_{RT},\ \mu_{RT}).SESS^{SB}_{Idle} \\ &+ (NHO_{2i-1,2i},\ P_C * \top).SESS^{SB}_{RT} \\ &+ (HHO\_blk,\ \top).SESS^{SB}_{Idle} \end{aligned}$$

The network selection for a new session request is still implemented in the component NS^{SB} by enabling corresponding activities.[3] The component NS^{SB} is defined as:

$$\begin{aligned} NS^{SB} &\stackrel{def}{=} (sess\_req^{C}_{RT},\ P_C * \top).NS^{SB} \\ &+ (sess\_req^{W}_{NRT},\ P_W * \top).NS^{SB} \end{aligned}$$

6.4 System Definition for Each Network Selection Strategy

The system definitions of the PEPA models of the above three strategies have the same structure: the components NS^G (NS^{WF}), and *SESS* cooperate with the component *MN* so as to control the session and network selection behaviour of the mobile node. The cooperation sets in the system definitions of the general and WLAN-first strategies are the same, while those for the service-based strategy are different from the others. The three PEPA models are defined as:

$$\begin{aligned} NSS^{G} &\stackrel{def}{=} SESS_{Idle} \underset{L_1}{\bowtie} MN_1^{Idle} \underset{L_2}{\bowtie} NS^{G}, \\ NSS^{WF} &\stackrel{def}{=} SESS_{Idle} \underset{L_1}{\bowtie} MN_1^{Idle} \underset{L_2}{\bowtie} NS^{WF}, \\ NSS^{SB} &\stackrel{def}{=} SESS^{SB}_{Idle} \underset{L_3}{\bowtie} MN_1^{Idle} \underset{L_4}{\bowtie} NS^{SB}, \end{aligned}$$

where the cooperation sets are

$$\begin{aligned} L_1 &= \{sess\_req_{NRT}, sess\_req_{RT}, sess\_req^{C}_{NRT}, sess\_req^{W}_{NRT}, \\ &\quad sess\_req^{C}_{RT}, sess\_req^{W}_{RT}, sess_{NRT}, sess_{RT}, HHO\_blk, VHO\_blk\}, \\ L_2 &= \{sess\_req^{C}_{NRT}, sess\_req^{W}_{NRT}, sess\_req^{C}_{RT}, sess\_req^{W}_{RT}, \\ &\quad NHO_{2i-1,2i}, VHO_{2i-1,2i}, VHO\_blk\}, \\ L_3 &= \{sess\_req_{NRT}, sess\_req_{RT}, sess\_req^{C}_{NRT}, sess\_req^{W}_{NRT}, \\ &\quad sess\_req^{C}_{RT}, sess\_req^{W}_{RT}, sess_{NRT}, sess_{RT}, HHO\_blk, VHO\_blk, \\ &\quad NHO_{2i-1,2i}, VHO_{2i-1,2i}\}, \\ L_4 &= \{sess\_req^{C}_{NRT}, sess\_req^{W}_{NRT}, sess\_req^{C}_{RT}, sess\_req^{W}_{RT}\}. \end{aligned}$$

We denote a system state of a PEPA model as $s_k^{A,B}$, where k, A and B represent the mobile node's phase of its mobility model, the RAN it is connected to, and the type of the session it is engaged in respectively. For example, $s_3^{C,RT}$ means the mobile node is in phase 3 and is connected to the 3GRAN for an RT session. Moreover, s_k^C and s_k^W are the unions of system states and they are defined as $s_k^C = s_k^{C,NRT} \cup s_k^{C,RT}$ and $s_k^W = s_k^{W,NRT} \cup s_k^{W,RT}$. Note that the system states of all the models are the *feasible* combinations of the state of each component of that model. For example, the model of the WLAN-first NSS does not have the states $s_{2l}^{C,RT}$ and $s_{2l}^{C,NRT}$, and the model of the service-based NSS does not have the states $s_{2l}^{W,RT}$ and $s_{2l}^{C,NRT}$, where $l = 1, 2, \cdots, N/2$.

3. In fact, the network selection for a new session request can also be implemented by modifying the idle state of the component *SESS*. The component NS^{SB} is deliberately used so that the model has the same structure as the other models.

6.5 Performance Measures

In this work, we investigate three performance measures, namely average throughput, RAN blocking probability and handover rate.

Average Throughput: The average throughput is defined as the mean data rate that can be achieved by a mobile node during its communication. To derive this measure, the first step is to obtain the percentages of time the mobile node spends using different RATs for different types of sessions. Therefore, four types of engaged times can be defined as follows:

$$\begin{aligned} T_{C,NRT} &= \sum_{i=1}^{N} \pi(s_i^{C,NRT}), \quad T_{C,RT} = \sum_{i=1}^{N} \pi(s_i^{C,RT}), \\ T_{W,NRT} &= \sum_{i=1}^{N/2} \pi(s_{2i}^{W,NRT}), \quad T_{W,RT} = \sum_{i=1}^{N/2} \pi(s_{2i}^{W,RT}), \end{aligned} \tag{2}$$

where $\pi(s_k^{A,B})$ is the equilibrium probability of system state $s_k^{A,B}$. Then the total percentage of time the mobile node is in the engaged states is:

$$T_{Engaged} = T_{C,NRT} + T_{C,RT} + T_{W,NRT} + T_{W,RT}. \tag{3}$$

Based on the above definitions, the average throughput is calculated as the weighted sum of the proportions of the different engaged times to the total engaged time, where the weights are the corresponding data rates of different RATs. That is:

$$\begin{aligned} THP =& D^C_{NRT} * \frac{T_{C,NRT}}{T_{Engaged}} + D^C_{RT} * \frac{T_{C,RT}}{T_{Engaged}} \\ &+ D^W_{NRT} * \frac{T_{W,NRT}}{T_{Engaged}} + D^W_{RT} * \frac{T_{W,RT}}{T_{Engaged}}, \end{aligned} \tag{4}$$

where D^C_{NRT}, D^C_{RT}, D^W_{NRT} and D^W_{RT} are the data rates that can be achieved by the mobile node when it uses the 3G RAT (3GRAT) and WLAN RAT (WRAT) for NRT and RT sessions respectively.

RAN Blocking Probability: Although the blocking probabilities of different RANs can be regarded as independent input parameters to the PEPA models, in this work an approach which utilises the PEPA models to derive the blocking probabilities is presented.

Handover Rate: The handover rate is defined as the mean number of handover attempts performed by the mobile node per unit time.

7 Derivation of RAN Blocking Probability and Handover Rate

For a certain type of NSS, the blocking probabilities of the 3GRAN and WRAN are derived from the interaction

between its PEPA model and a resource consumption model corresponding to that type of NSS. In this procedure, the horizontal and vertical handover rates of the mobile node are obtained at the same time. In the following subsections, the mathematical expressions of the RAN blocking probability and handover rate are presented first, followed by an iterative method to derive them.

7.1 RAN Blocking Probability

To derive blocking probabilities of RANs, a two-dimensional continuous-time Markov chain (2D-CTMC) is used to model the resource consumption of a 3G-WLAN interworking cell. The state of the 2D-CTMC is denoted by two nonnegative integers (c, w), where c and w are the numbers of engaged users in the 3GRAN and WRAN respectively. For WLAN cells which overlap with two adjacent cellular cells, their resources are assumed to be shared by both 3G-WLAN interworking cells. That is, the changes in the number of users of these WLAN cells are reflected in the changes of the states of their spanning 3G-WLAN interworking cells. As shown in Fig. 4, there are five types of events that change the state of the 2D-CTMC and they are described as follows:

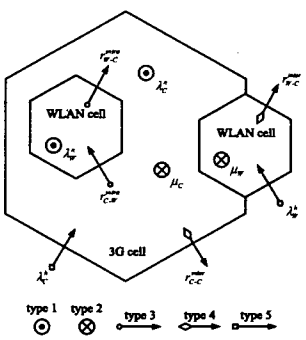


Fig. 4. Five types of events that change the state of the 2D-CTMC

- **type 1**: New sessions requests are generated in the 3GRAN and the WRAN, and their rates are denoted as λ_C^n and λ_W^n respectively.
- **type 2**: Sessions are completed and resources in the 3GRAN and the WRAN are released, and their rates are denoted as μ_C and μ_W respectively.
- **type 3**: Sessions are *internally* handed over *between* the 3GRAN and the WRAN. Their rates are denoted as r_{C-W}^{intra} and r_{W-C}^{intra} respectively.
- **type 4**: Sessions are *externally* handed over *out of* the 3GRAN and the WRAN. Their rates are denoted as r_{C-C}^{inter} and r_{W-C}^{inter} respectively.
- **type 5**: Sessions are *externally* handed over *into* the 3GRAN and WRAN and their rates are denoted as λ_C^h and λ_W^h respectively.

According to the events described above, the state-transition diagram of the 2D-CTMC can be generated. Fig. 5 shows the outward transitions of a non-boundary state (c, w) of the 2D-CTMC. The whole state-transition diagram of the 2D-CTMC can be constructed straightforwardly.

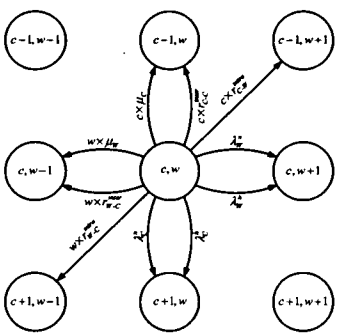


Fig. 5. Outward transitions from a non-marginal state of the 2D-CTMC

For each type of NSS, the rates of the transitions corresponding to the five types of events are calculated as follows:

- **type 1**: Assume that the mobile nodes in the 3G-WLAN interworking cell are uniformly distributed and let A_C and A_W denote the coverage percentage of 3G area and the 3G-WLAN area respectively. For the general and WLAN-first strategies, λ_C^n and λ_W^n are calculated as:

$$\lambda_C^n = A_C * \Lambda^n + P_C * A_W * \Lambda^n, \quad \lambda_W^n = P_W * A_W * \Lambda^n, \quad (5)$$

and for the service-based strategy they are calculated as:

$$\lambda_C^n = A_C * \Lambda^n + P_{RT} * A_W * \Lambda^n, \quad \lambda_W^n = P_{NRT} * A_W * \Lambda^n, \quad (6)$$

where Λ^n is the arrival rate of the new session requests of the whole 3G-WLAN interwork cell.

- **type 2**: For the general and WLAN-first strategies, the resources of the 3GRAN and the WRAN can be used by both NRT and RT users. Therefore in both RANs the probabilities that a session is NRT or RT are P_{NRT} or P_{RT} respectively, and the resources holding time in the 3GRAN $(1/\mu_C)$ and the WRAN $(1/\mu_W)$ are calculated as:

$$\frac{1}{\mu_C} = \frac{P_{NRT}}{R_{NRT}^C * \mu_{NRT}} + \frac{P_{RT}}{\mu_{RT}}, \quad \frac{1}{\mu_W} = \frac{P_{NRT}}{R_{NRT}^W * \mu_{NRT}} + \frac{P_{RT}}{\mu_{RT}}. \quad (7)$$

For the service-based strategy, resource consumption is a bit more complex. The resources of the WRAN can only be used by NRT users in the 3G-WLAN area, whereas the resources of the 3GRAN can be used by RT users in the 3G-WLAN area and all the users in the 3G area. To simplify the analysis, it is assumed that the probabilities that a session in the 3GRAN is NRT and RT are still P_{NRT} and P_{RT} respectively. Accordingly, $1/\mu_C$

and $1/\mu_W$ are calculated as:

$$\frac{1}{\mu_C} \approx \frac{P_{NRT}}{R^C_{NRT} * \mu_{NRT}} + \frac{P_{RT}}{\mu_{RT}}, \quad \frac{1}{\mu_W} = \frac{1}{R^W_{NRT} * \mu_{NRT}}. \tag{8}$$

- **type 3 and 4**: r^{intra}_{C-W}, r^{intra}_{W-C}, r^{inter}_{C-C} and r^{inter}_{W-C} are derived by the approach which will be discussed in Section 7.2.
- **type 5**: λ^h_C and λ^h_W are regarded as input parameters.

Once the transition rates are obtained, the generator matrix of the 2D-CTMC can be generated and its equilibrium probability vector can be derived. Then the blocking probabilities of the 3GRAN and the WRAN are calculated as:

$$P^C_B = \sum_{\substack{c=N_C \\ 0\leqslant w\leqslant N_W}} p(c,w), \quad P^W_B = \sum_{\substack{w=N_W \\ 0\leqslant c\leqslant N_C}} p(c,w), \tag{9}$$

where $p(c,w)$ is the equilibrium probability of the state (c,w) of the 2D-CTMC, and N_C and N_W are the maximum number of 3G and WLAN users that can be supported in a 3G-WLAN interworking cell respectively.[4] Note that a different 2D-CTMC is required for each NSS since their generator matrices are different. Eq. (9) are the general expressions for all types of selection strategies.

7.2 Handover Rate

Since the handover rate is defined as the mean number of handover *attempts* performed by the mobile node per unit time, it is actually the throughput of the handover activity. Therefore, the handover rate can be calculated using MRS and Eq. (1), where the rewards are equal to the *activity rate* of that type of handover and they are associated with the system states in which that type of handover is enabled. For example, the activity rate of the internal vertical handover from WRAN to 3GRAN is $a_{2i} * v_{2i}$, and the mobile node performs this type of handover in the system states s^W_{2i}.

Four types of handovers are expressed by the mobility model and their corresponding handover rates are calculated as follows:

$$\begin{aligned}
r^{inter}_{C-C} &= \sum_{i=1}^{N/2} \left(b_{2i-1} * v_{2i-1} * \pi(s^C_{2i-1}) + b_{2i} * v_{2i} * \pi(s^C_{2i}) \right), \\
r^{inter}_{W-C} &= \sum_{i=1}^{N/2} b_{2i} * v_{2i} * \pi(s^W_{2i}), \\
r^{intra}_{C-W} &= \sum_{i=1}^{N/2} P_W * a_{2i-1} * v_{2i-1} * \pi(s^C_{2i-1}), \\
r^{intra}_{W-C} &= \sum_{i=1}^{N/2} a_{2i} * v_{2i} * \pi(s^W_{2i}),
\end{aligned} \tag{10}$$

where $\pi(s^C_{2i-1})$, $\pi(s^C_{2i})$ and $\pi(s^W_{2i})$ are the equilibrium probabilities of the corresponding system states of the PEPA models.

4. The capacity of the 3G-WLAN interworking cell is a complex topic and depends on the user data rate requirements, signal-to-noise ratio, etc [21]. The static capacity is used in order to simply the analysis.

7.3 An Iterative Method to Derive RAN Blocking Probability and Handover Rate

As discussed above, the derivation of RAN blocking probabilities $\boldsymbol{P_B} = [P^C_B, P^W_B]$ requires the handover rates $\boldsymbol{R_H} = [r^{intra}_{C-W}, r^{intra}_{W-C}, r^{inter}_{C-C}, r^{inter}_{W-C}]$ be derived from the PEPA models to calculate $\boldsymbol{p}(c,w)$ of the 2D-CTMC model. On the other hand, $\boldsymbol{R_H}$ are calculated from $\pi(s^{A,B}_k)$ of the PEPA models, which requires $\boldsymbol{P_B}$ be derived from the 2D-CTMC model. Therefore, for each type of NSS, its corresponding 2D-CTMC and PEPA models form a closed loop by exchanging $\boldsymbol{P_B}$ and $\boldsymbol{R_H}$, i.e.,

$$\begin{aligned}
F_{CTMC}(\boldsymbol{R_H}): \quad \boldsymbol{R_H} &\xrightarrow{f_{CTMC}} \boldsymbol{p}(c,w) \xrightarrow{(9)} \boldsymbol{P_B}, \\
F_{PEPA}(\boldsymbol{P_B}): \quad \boldsymbol{P_B} &\xrightarrow{f_{PEPA}} \pi(s^{A,B}_k) \xrightarrow{(10)} \boldsymbol{R_H},
\end{aligned}$$

where f_{CTMC} and f_{PEPA} denote the mathematical computation to obtain the equilibrium probability vectors of the 2D-CTMC and the PEPA models respectively. The two procedures to derive $\boldsymbol{P_B}$ and $\boldsymbol{R_H}$ are denoted as $F_{CTMC}(\boldsymbol{R_H})$ and $F_{PEPA}(\boldsymbol{P_B})$ respectively. To solve this implicit problem, an iterative method is designed and it is described in Algorithm 1.

Algorithm 1: Derivation of $\boldsymbol{P_B}$ and $\boldsymbol{R_H}$

Input: all the required parameters
Output: $\boldsymbol{P_B} = [P^C_B, P^W_B]$, $\boldsymbol{R_H} = [r^{intra}_{C-W}, r^{intra}_{W-C}, r^{inter}_{C-C}, r^{inter}_{W-C}]$.

```
conv = 0 ;                          /* convergence tag */
ε = 10^-30 ;                        /* convergence criteria */
i = 1 ;                             /* iteration counter */
iter = 100 ;                        /* maximum iteration number */
P_B^(1) = [0,0] ;                   /* initialise P_B^(1) */
R_H^(1) = F_PEPA(P_B^(1)) ;         /* calculate R_H^(1) using P_B^(1) */
while (i ≤ iter) or (conv = 0) do
    P_B^(i+1) = F_CTMC(R_H^(i)) ;   /* update P_B^(i+1) using R_H^(i) */
    err = max{P_B^(i+1) − P_B^(i)} ; /* calculate difference err */
    if (err > ε) then
        R_H^(i+1) = F_PEPA(P_B^(i+1)) ; /* calculate R_H^(i+1) using P_B^(i+1) */
        R_H = R_H^(i+1) ;           /* update R_H */
        P_B = P_B^(i+1) ;           /* update P_B */
        i = i + 1;
    else
        conv = 1 ;                  /* set the convergence tag */
        R_H = R_H^(i) ;             /* update R_H */
        P_B = P_B^(i) ;             /* update P_B */
    end
end
return P_B, R_H;
```

The convergence speed of the above iterative method is dependent on the parameter setting but very fast. For example, Table 1 lists the numbers of iterations executed to derive results from each model for 10 increasing session durations. The other parameters are set according to traffic pattern 2, which will be described later.

TABLE 1
Numbers of iterations executed to derive results from each model for 10 increasing session durations listed in TABLE 3

Model	Numbers of iterations
Random	[2, 2, 3, 4, 5, 7, 9, 11, 11, 13]
RRSS	[2, 2, 3, 4, 5, 7, 8, 11, 12, 13]
WLAN-first	[2, 2, 3, 4, 5, 6, 8, 10, 12, 13]
Service-based	[2, 2, 3, 4, 5, 6, 8, 10, 12, 13]

8 Performance Evaluation

In this work, four strategies are investigated, namely: random, RRSS, WLAN-first and service-based. The effect of different mobility and traffic patterns of a mobile node on the average throughput, RAN blocking probability and handover rate are investigated.

8.1 Parameter Settings

TABLE 2
Parameter settings of the 2D-CTMC and PEPA models

Branching probabilities			
a_1	0.7	b_1	0.3
a_2	0.5	b_2	0.5
a_3	0.3	b_3	0.7
a_4	0	b_4	1
Data rates and data factors			
D^C_{NRT}	2 Mbps [22]	D^W_{NRT}	6 Mbps [21]
D^C_{RT}	2 Mbps	D^W_{RT}	2 Mbps
R^C_{NRT}	1	R^W_{NRT}	3
Capacities, coverage percentages and session arrival rates			
N_C	50	N_W	30
A_C	0.95	A_W	0.05
λ^h_C	1/20	λ^h_W	1/20
Λ^n	1/10		
Network selection probabilities			
Random	$P_C = 0.5$	$P_W = 0.5$	
RRSS	$P_C = 0.4$	$P_W = 0.6$ [10]	
WLAN-first	$P_C = 0$	$P_W = 1$	
Service-based	$P_C = 1$ (for RT session)		
	$P_W = 1$ (for NRT session)		

Before the evaluation, we need to set up the parameters used in the models. Table 2 lists the settings of the parameters used in the 2D-CTMC and PEPA models. They are divided up into four groups: (a) the phase branching probabilities of the mobility model which is assumed to have four phases; (b) the data rates of the 3GRAT and WRAT for the NRT and RT sessions, and the corrresponding data rate factors; (c) the capacity, coverage percentages of the 3GRAN and the WRAN[5], and the session arrival rates of the 3G-WLAN interworking cell; (d) the network selection probabilities of different strategies.

TABLE 3
Activity rates of the PEPA models

Rate	Descriptions	Value (second^{-1})
λ	session arrival rate	180^{-1}
υ_1, υ_3	mobility rate of phase 1 and 3	600^{-1}
υ_2, υ_4	mobility rate of phase 2 and 4	474^{-1}, 1200^{-1}
μ_{RT}	rate of RT sessions	$[60 : 60 : 600]^{-1}$
μ_{NRT}	rate of NRT sessions	$[30 : 30 : 300]^{-1}$
P_{NRT}	NRT session generation probability	0.3, 0.7

The activity rates used in the PEPA models of different NSSs are listed in Table 3. The duration of the NRT session $1/\mu_{NRT}$ and that of the RT session $1/\mu_{RT}$ are the control parameters in all the evaluation. The duration of the NRT session is measured at 2 Mbps and thus it reflects the traffic volume of the NRT session. The session arrival interval of a mobile node is set to 180 seconds. The mean sojourn time of the mobile node in the 3G area ($1/\upsilon_1$ and $1/\upsilon_3$) is assumed to be the same and is set to 600 seconds. Two mobility patterns and two traffic patterns are investigated, and they are controlled by the mean sojourn time of the mobile node in the 3G-WLAN area ($1/\upsilon_2$ and $1/\upsilon_4$) and the proportion of the NRT session generated by the mobile node (P_{NRT}).

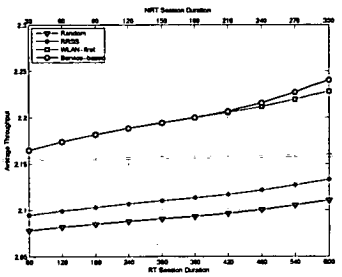
8.2 Effects of Mobility Pattern

The mobility pattern of the mobile node is controlled by its mean sojourn time in the 3G-WLAN area. Two scenarios are considered in the evaluation.

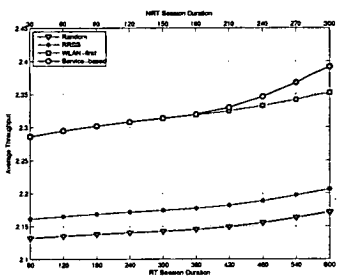
1) In the first scenario, the fluid flow movement model [23] is employed. Based on the formulas developed for the model, and with the assumptions on cell shape and radius, the mean sojourn time of a mobile node in the 3G area and the 3G-WLAN area is 38/30. Then if υ_1 and υ_3 are 600^{-1} then υ_2 and υ_4 are about 474^{-1}.
2) In the second scenario, the mobile node spends a longer time in the 3G-WLAN area and υ_2 and υ_4 are set to 1/1200.

Note that the number of phases, the sojourn times and the branching probabilities of the mobility model used in the evaluation are contrived, more practical values can be estimated from field data using algorithms studied in [16], [18]. NRT and RT sessions are generated at equal probabilities, i.e., $P_{NRT} = P_{RT} = 0.5$. The other

5. We assume there are 5 circular WLAN cells in a circular 3G-WLAN interworking cell, and their radii are set to 100m and 1000m respectively so that the data rates of 6 Mbps and 2 Mbps can be achieved.

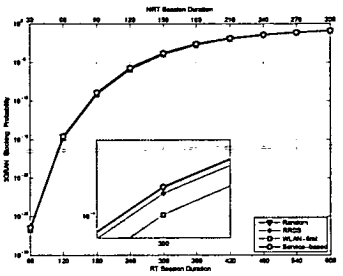


(a) Mobility pattern 1

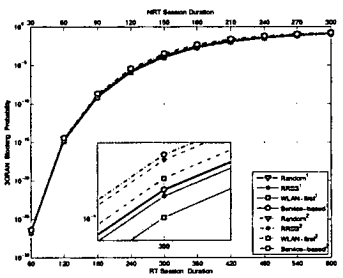


(b) Mobility pattern 2

Fig. 6. The effect of mobility pattern on average throughput



(a) Mobility pattern 1



(b) Mobility pattern 1 and Mobility pattern 2

Fig. 7. The effect of mobility pattern on 3GRAN blocking probability

parameters and activity rates are the same as listed in Table 2 and Table 3. In all the figures, the investigated performance measures are plotted against the session duration. Since there are two types of sessions, two x-axes are used where the top x-axis is the NRT session duration and the bottom x-axis is the RT session duration.

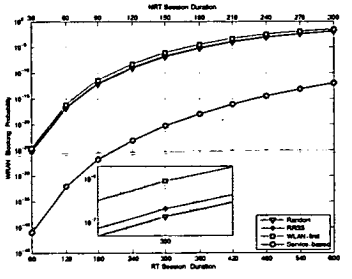
8.2.1 Average Throughput

Fig. 6 shows the average throughput achieved by the mobile node with different mobility patterns. According to Eq. (4), if $D_W^{NRT} = 2$, then the average throughput is always 2 Mbps. Therefore, the average throughput mostly depends on how much time an NRT session uses the WRAT. Fig. 6 indicates that a higher average throughput can be gained by using the strategy with a larger WRAN selection probability, or by simply staying in the 3G-WLAN area for a longer time. Moreover, a longer NRT session duration also results in a higher average throughput, because the mobile node has more opportunity to use the WRAT. An interesting observation is that when NRT session duration is less than 210 seconds, the WLAN-first and service-based strategies have almost the same performance on average throughput but for longer session durations, more of an NRT session in the service-based strategy is spent on the WRAT which results in a clear improvement on the average throughput. This result suggests that the service-based strategy can make the best use of the high data rate of the WRAN, especially for long NRT sessions.

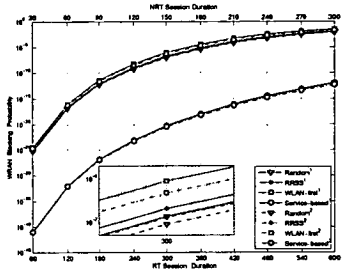
8.2.2 RAN Blocking Probability

Fig. 7 shows the blocking probability of the 3GRAN experienced by the mobile node with different mobility patterns. Note that the y-axis is a logarithmic scale. Since the 2D-CTMC resource consumption model does not include any handover prioritised scheme, the derived blocking probability is for both new and handover session requests. The results indicate that the blocking probability of the 3GRAN mostly depends on its traffic load, which is decided by *how frequently* it is chosen for a session and *how long* the session engages the resources. From Fig. 7(a), it can be observed that the service-based and random strategies are very close and also have the highest blocking probabilities. This is because given $P_{NRT} = P_{RT} = 0.5$, the random and service-based strategies have the same and also the highest probability of using 3GRAN. This can be explained as follows: in the service-based strategy, although only RT sessions choose the 3GRAN, the probability that a session is RT is 0.5; whereas in the random strategy, both types of session can choose the 3GRAN with the probability of 0.5. However, since the approximation made in Eq. (8) is an underestimate[6], the service-based strategy should

6. This is because the RT session, which has a longer duration, should have a proportion larger than P_{RT}.

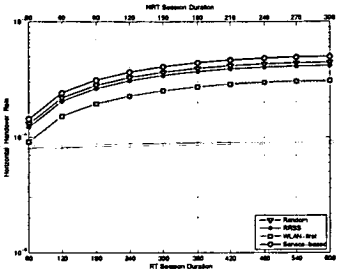


(a) Mobility pattern 1

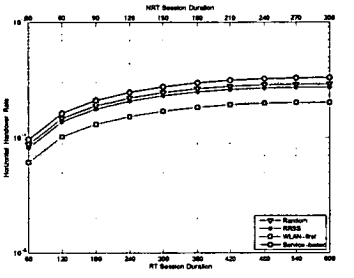


(b) Mobility pattern 1 and Mobility pattern 2

Fig. 8. The effect of mobility pattern on WRAN blocking probability



(a) Mobility pattern 1



(b) Mobility pattern 2

Fig. 9. The effect of mobility pattern on horizontal handover rate

have a clearly higher 3GRAN blocking probability than the random strategy. The RRSS strategy ranks third as all the sessions choose the 3GRAN with the probability of 0.4 and the WLAN-first strategy has the lowest 3GRAN blocking probability. In Fig. 7(b), the effect of the mobile node's mobility is shown. For all of the strategies, a longer stay in the 3G-WLAN area results in a higher 3GRAN blocking probability.

Similarly, the blocking probabilities of the WRAN also depends on its traffic load and the differences between the different strategies are more obvious as shown in Fig. 8. It can be observed that the WRAN blocking probability of the service-based strategy is much smaller than the others. This is because only NRT sessions are allowed to use the WRAT and more importantly they engage the WRAN resources for a shorter time than the RT sessions. Moreover, the mobility of the mobile node has the opposite effect on the WRAN blocking probability to that on the 3GRAN, i.e., a longer stay in the 3G-WLAN area results in a lower WRAN blocking probability.

8.2.3 Handover Rate

Fig. 9 shows the horizontal handover rate performed by the mobile node with different mobility patterns. A horizontal handover happens when the engaged mobile node moves across adjacent 3G-WLAN interworking cells. According to Eq. (10), the horizontal handover rate depends on the probability that the mobile node is using the 3GRAT at the time it moves out of its current 3G-WLAN interworking cell. The results indicate that a mobile node using the service-based strategy is the most likely to perform a horizontal handover. This is because the service-based strategy makes the mobile node spend longer in the 3GRAN than the other strategies as explained in Section 8.2.2. The random strategy ranks the second and the WLAN-first strategy has the smallest horizontal handover rate as it uses the 3GRAT less than the other strategies. Given a certain mobility pattern, a longer session duration means the mobile node is more likely to hand over during a session and thus results in a higher handover rate. Moreover, given a certain session duration, mobility can reduce the handover rate for all of the strategies, as shown in Fig. 9(b).

The vertical handover is defined as the handover between different RATs. This rate depends on the probability that the mobile node is using the WRAT at the time it moves out of the 3G-WLAN area, and the probability of choosing the WRAN when it moves into the 3G-WLAN area. Therefore, as shown in Fig. 10, the WLAN-first strategy experiences the most frequent vertical handover whereas the service-based strategy has the lowest vertical handover rate. The effect of the session duration and mobility on the vertical handover rate are the same as those on the horizontal handover rate.

Appendix A
Publications

The author of this thesis has the following accepted or submitted publications during the course of his Ph.D. research:

A.1 Journal Papers

- Hao Wang, David I. Laurenson, and Jane Hillston, "A General Performance Evaluation Framework for Network Selection Strategies in 3G-WLAN Interworking Networks," submitted to *IEEE Transaction on Mobile Computing*. (on page 145)

- Hao Wang, David I. Laurenson, and Jane Hillston, "A Reservation Optimised Advance Resource Reservation Scheme for Deploying RSVP in Mobile Environments," *Springer Wireless Personal Communications*, published online first, 2009. (on page 161)

A.2 Conference Papers

- Hao Wang, David I. Laurenson, and Jane Hillston, "Evaluation of RSVP and Mobility-Aware RSVP Using Performance Evaluation Process Algebra," in *Proc. IEEE International Conference on Communications*, 2008, pp. 192–197. (on page 187)

- Hao Wang, David I. Laurenson, and Jane Hillston, "An SMR based advance resource reservation scheme for combined mobility and QoS Provisioning," in *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, 2008, pp. 1–5. (on page 192)

- Hao Wang, David I. Laurenson, and Jane Hillston, "PEPA Analysis of MAP Effects in Hierarchical Mobile IPv6," in *Proc. IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, 2007, pp. 337–342. (on page 197)

The original publications are included in the following pages.

A General Performance Evaluation Framework for Network Selection Strategies in 3G-WLAN Interworking Networks

Hao Wang, *Student Member, IEEE*, David I. Laurenson, *Member, IEEE*, and Jane Hillston, *Member, IEEE*

Abstract—In this work we investigate a general performance evaluation framework for network selection strategies (NSSs) that are used in 3G-WLAN interworking networks. Instead of simulation, this framework is based on models of NSSs and is constructed using a stochastic process algebra, named Performance Evaluation Process Algebra (PEPA). It captures the traffic and mobility characteristics of mobile nodes in 3G-WLAN interworking networks and has a good expression of the behaviour of the mobile nodes using different NSSs. Commonly used NSSs are evaluated from the perspectives of average throughput, handover rate, and network blocking probability. Results of the evaluation explore the effect of these NSSs on both mobile nodes and networks, as well as their characteristics in different mobility and traffic scenarios.

Index Terms—3G, WLAN, network selection, performance evaluation, Markov chain, process algebra.

1 Introduction

With the rapid development of various wireless communication technologies, the main trend for future wireless communications is a shift from voice and text based services provided by early cellular networks to multimedia-based services provided by multiple and heterogeneous wireless networks. These multimedia applications require different bearer services and there is no single technology that can simultaneously provide low latency, cheap cost, high data rate and high mobility to mobile users [1]. Therefore, next-generation wireless communications focuses on the integration of existing cellular and other wireless networks. Especially, owing to the recent evolution and successful deployment of the wireless local area network (WLAN), there has been a demand for integrating WLAN with the third-generation (3G) cellular network, i.e. 3G-WLAN interworking. For example, the 3rd generation partnership project (3GPP), an international organisation which produces technical specifications and reports for 3G wireless wide area networks (WWANs), has developed an architecture for the integration of the 3GPP cellular network and WLAN with the intention to extend 3GPP services to the WLAN access environment [2].

In a 3G-WLAN interworking network, a mobile node may perform vertical handovers (VHOs) [3] when it moves across overlaid 3G and WLAN radio access networks (RANs). During the handover, the mobile node may use its network selection strategy (NSS) [4] to select a certain type of radio access technology (RAT) to continue its communication. Various handover criteria can be taken into account when making a vertical handover decision, e.g., cost of services, network and mobile node conditions, and user preference, etc [5]. NSSs have been designed to abstract the process of selecting an appropriate network as mathematical problems which assess alternative networks in terms of different criteria. For a general review of the proposed strategies and their details, refer to [6], [7] and the references therein. Since NSSs control the session behaviour of mobile nodes, it is important and meaningful to investigate how they affect the performance of user applications at the mobile nodes and the utilisation of the RANs. Moreover, it would be better if the assessment can be conducted using a formal performance modelling technique rather than system-level simulations.

In this work, we investigate a general performance evaluation framework for NSSs in 3G-WLAN interworking networks. The framework is based on the performance models constructed using a formal performance modelling technique, named Performance Evaluation Process Algebra (PEPA). PEPA is chosen because its behavioural and compositional system description capability can efficiently reduce the modelling and evaluation difficulty of complex systems, and therefore facilitate the whole evaluation process. This framework captures the traffic and mobility characteristics of mobile nodes in 3G-WLAN interworking networks and has a good expression of the behaviour of the mobile nodes using different NSSs. From the framework, important performance measures including average throughput, RAN blocking probability and handover rate are derived.

The rest of this paper is structured as follows. Section 2 gives a brief review of the previous work on evaluating NSSs used in 3G-WLAN interworking networks. A traffic model of multimedia communications and a

This research has been funded by the Industrial Companies who are Members of Mobile VCE, with additional financial support from the UK Governments Technology Strategy Board (previously DTI).

mobility model suitable for 3G-WLAN interworking networks which are used in the framework are presented in Section 3 and Section 4 respectively. In Section 5, a short introduction to the PEPA formalism is given and in Section 6 PEPA models of different NSSs are described. In Section 7, an iterative method that interlinks the PEPA models and a network resource consumption model in order to derive the RAN blocking probability and handover rate is described. The performance of different NSSs are evaluated in Section 8 and in Section 9 the evaluation results are discussed.

2 Related Work and Contributions

In previous studies on the design and evaluation of NSSs in 3G-WLAN interworking networks, the selection strategies are first formulated as policy functions which take network and mobile node conditions as the input parameters and then output the preferred RAN. Then comparisons are made between the results of different strategies, or between the results of the same strategy under different scenarios, in terms of throughput, connection cost, power consumption, etc. Typical examples of this can be found in [6], [8], [9]. There are also studies which aim to integrate resource management schemes into the NSSs. In [10], network selection takes account of the traffic loads of different RANs in order to achieve a balance between the RANs. In [11], [12], the mobility, traffic and location information of the mobile nodes are considered, and network selection is implemented in a centralised way to manage the resources of the RANs more efficiently. An analytical model for evaluating different NSSs is investigated in [13]. Each strategy is formulated as a mapping function between the total session arrival rates of the whole 3G-WLAN interworking network and the session arrival rates into different RANs. However, this work only considers network selection for new session requests and node mobility is not taken into account.

One major limitation of the above studies is that the evaluation is usually carried out in a restricted way. That is, different strategies are compared at the policy function level without considering the mobile node's session and handover behaviour, whereas evaluation with consideration of traffic and mobility only focuses on a certain type of strategy. However, to make a fair comparison, different strategies should be evaluated using the same general framework with as few restrictions as possible. To the best of our knowledge, there are no such framework for evaluating NSSs in the published literature, and this work is thus motivated by investigating such a general performance evaluation framework. The contributions of this work are summarised as follows:

- With the help of the behavioural and compositional system description capability of the PEPA language, in the framework, different NSSs are modelled in terms of the behaviour of a mobile node. To make the description more accurate, a traffic model which represents features of multimedia-based services, and a mobility model which captures movement characteristics of a mobile node, are employed.
- NSSs for both new and handover sessions are considered in the framework, and they are embedded in the form of network selection probabilities. This level of abstraction provides the framework the flexibility to express any type of NSS given that its probabilities of choosing RANs can be determined.
- The generality of the framework is also retained by having an interface to the model capturing RAN resource consumption, in the form of the network blocking probability. In this way, the framework is independent on the call admission control (CAC) and resource management schemes used in the 3G-WLAN interworking networks.
- Moreover, a novel iterative algorithm is proposed to derive network blocking probability from this framework when it is not at the modellers' disposal. This algorithm links models of RAN resource consumption and models of different NSSs through interchanging necessary parameters between them. The convergence speed of the algorithm is fast, and performance measures are direct results of it.

3 Traffic Model

The traffic of a mobile node is modelled at the session level in the framework, which includes two parameters: session arrival rate and session duration. Due to the multi-service characteristic of next-generation communications, the mobile node requires different types of services that have different durations. This combination of various services results in a large variance in the observed user session duration. Moreover, the field data suggests that the statistical session duration in the Internet has a coefficient of variation (CoV) larger than one [14], [15]. To capture this traffic feature, we use the hyper-exponential distribution to model the session duration of a mobile node.

A K-phase ($K > 1$) hyper-exponential distribution is composed of K exponentially distributed phases in parallel and always has its CoV larger than one. To simplify the traffic model, a two-phase hyper-exponential distribution is used, where one phase represents the non-real time (NRT) sessions and the other represents the real-time (RT) sessions. The NRT sessions generally include Web surfing and file transfer, and the RT sessions generally include voice and video streaming. As for the session arrival rate, the general consensus that the session arrival is a Poisson process is followed.

On the basis of the above assumptions, the traffic model can be constructed as a combination of two ON-OFF sources. As shown in Fig 1, a session arrives at the rate of λ and it can be either an RT session with probability P_{RT} or an NRT session with probability $P_{NRT} = 1 - P_{RT}$. The mean durations of the RT and NRT sessions are $1/\mu_{RT}$ and $1/\mu_{NRT}$ respectively.

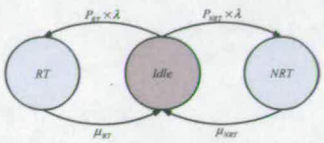


Fig. 1. A traffic model with two ON-OFF sources

4 Mobility Model

In a 3G-WLAN interworking network, a cellular cell is usually called a *3G-WLAN interworking cell*, and is generally overlaid with one or more WLAN cells. Therefore, the mobility model of a mobile node should characterise its residence times not only in the whole 3G-WLAN interworking cell but also in the different RAT areas within the 3G-WLAN interworking cell. To this end, the mobility of the mobile node in the 3G-WLAN interworking network is modelled as a continuous-time Markov chain (CTMC) which has the *Coxian* structure.

A K-phase Coxian structure is composed of a series of K exponentially distributed phases (or states) and an absorbing phase. Phase i either enters the phase $i+1$ or enters the absorbing phase with pre-defined probabilities. Phase K enters the absorbing phase with the probability of 1. A Coxian distribution is then defined as the distribution from the first phase until absorption, and it has an important property that it can arbitrarily closely approximate any probability distribution [16]. By letting the phases represent the position of the mobile node in a 3G-WLAN interworking cell in terms of the RAT area, the transitions of the Coxian structure can capture various tracks that the mobile node may traverse in a 3G-WLAN interworking cell [17].

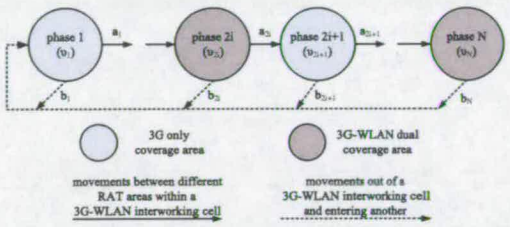


Fig. 2. A mobility model with the Coxian structure

Since the ordinary Coxian structure contains an absorbing state whereas the focus of this work is on the steady state performance of a mobile node, the absorbing state is omitted in the mobility model. The mobility model with the Coxian structure is assumed to have an even number (N) of phases, and it is shown in Fig. 2. The odd phases $2i-1$ and the even phases $2i$ represent the mobile node being in the *3G only coverage areas* and in the *3G-WLAN dual coverage areas* respectively, where $i = 1, 2, \ldots, N/2$. Without ambiguity, the above two types of area are called a *3G area* and a *3G-WLAN area* respectively. Two important assumptions are made in this work: Firstly, WLAN cells are assumed not to overlap with each other. Secondly, a WLAN cell which overlaps with two adjacent cellular cells is considered to belong to *both* cellular cells. With the second assumption, the starting point of the track of the mobile node in a 3G-WLAN interworking cell is always the 3G area.

The state transition diagram shown in Figure 2 captures the movement of a mobile node within and across the 3G-WLAN interworking cells: Movements across different RAT areas *within* a 3G-WLAN interworking cell are captured by the transitions between neighbouring phases. Movements *out of* the 3G-WLAN interworking cell are captured by the transitions from interim phases back to the first phase, and mean the mobile node enters the 3G area of *another new* 3G-WLAN interworking cell.

Owing to the Coxian structure of the mobility model, the sojourn time of the mobile node in the 3G-WLAN interworking cell, i.e., the time spent by the mobile node traversing a series of phases until going back to the first phase, still follows the Coxian distribution. In the mobility model, each phase of the mobility model, e.g. phase k, is assumed to be exponentially distributed with rate v_k, and the probabilities of branching to the next and the first phases are a_k and b_k respectively. The number of phases, the rate and the branching probabilities of each phase can be estimated from field data using algorithms presented in [16], [18].

5 Performance Evaluation Process Algebra

In this section, we give a short introduction to the PEPA language which is adopted as the modelling technique in this work.

Process algebras have been designed as formal description techniques for concurrent systems — systems that consist of subsystems interacting with each other. In process algebras, a *process* or an *agent* can perform *actions*, and a system is modelled from the perspective of its *behaviour* as interactions between processes. PEPA [19] is a stochastic extension of classical process algebras. It is important to emphasise that PEPA is not a Monte-Carlo evaluation approach but an advanced mathematical modelling technique that is ideal for performance modelling of large and complex resource-sharing systems.

5.1 Syntax of PEPA

PEPA is a compositional approach that decomposes a system into subsystems that are smaller and more easily modelled. In PEPA a system is usually modelled to be composed of a group of *components* that engage in *activities*. This abstract description of a system is similar to the design process of a system and facilitates model construction. Generally, components model the physical or logical elements of a system and activities characterise

the behaviour of these components. For example, a printing system can be considered to consist of a *Queue* component which buffers jobs and a *Printer* component which prints jobs.

Each activity a in PEPA is defined as a pair (α, r) — the action type α, which can be regarded as the name of the activity, and the activity rate r, which is the parameter of an exponentially distributed random variable that specifies the duration of the activity. If a component P behaves as P' after completing activity a, then we can regard this behaviour as a component changing from state P to state P', through transition (α, r).

The PEPA formalism provides a small set of operators which are able to express the individual activities of components as well as the interactions between them. Here we only present the operators that are used in this work. For more details about PEPA operators, see [19].

Prefix: $(\alpha, r).P$

This component has a designated first activity which is of action type (or name) α and has a duration that is exponentially distributed with rate r, which gives a mean time of $1/r$. After completing this activity, the component $(\alpha, r).P$ behaves as P. For example, the *Printer* component in the above example can print a job and then suspend for a while before it is ready to take the next job. This behaviour can be expressed as:

$$Printer \stackrel{def}{=} (print, r_1).(suspend, r_2).Printer$$

Choice: $P + Q$

This component may either behave as P or Q. All the enabled activities in P and Q are enabled in this component and compete with each other. The first activity to be completed will be an activity of P or Q and this will distinguish which component wins the race. When the first activity is completed, all the other activities will be abandoned. For example, let the component $Queue_i$ denote the behaviour of the *Queue* component when there are i jobs in the queue. It can either allow another job to arrive (when the queue is not full) or have one of its jobs printed (when the queue is not empty). The *Queue* component can then be defined as $(i = 1, 2, \cdots, N-1)$:

$$\begin{aligned} Queue_0 &\stackrel{def}{=} (arrive, r_3).Queue_1 \\ Queue_i &\stackrel{def}{=} (arrive, r_3).Queue_{i+1} \\ &\quad + (print, \top).Queue_{i-1} \\ Queue_N &\stackrel{def}{=} (print, \top).Queue_{N-1} \end{aligned}$$

where N is the maximum size of the queue. The symbol "$\top$" means the rate of the activity is outside control of the component. In this example, the *Queue* component is *passive* with respect to the activity *print* since it cannot influence the rate at which jobs are printed.

Cooperation: $P \underset{L}{\bowtie} Q$

This component represents the interaction between P and Q. The set L is called the *cooperation set* and denotes a set of action types that must be carried out by P and Q together. For all activities whose action type is included in set L, P and Q must cooperate to complete it. However, other activities of P and Q which have types that are not included in set L will proceed independently. The rate of the *shared* activity is determined by the rate of the slower participant and is the smaller of the two rates. When the cooperation set L is empty, the two components proceed concurrently without any interaction between them. A shorthand notation $P \| Q$ is used to represent $P \underset{\varnothing}{\bowtie} Q$. For example, the printing system may have two parallel queues that share only one printer, and each *Queue* component cooperates with the *Printer* component on the activity *print* individually. This can be expressed as:

$$(Queue_0 \| Queue_0) \underset{\{print\}}{\bowtie} Printer$$

Constant: $P \stackrel{def}{=} Q$

The constant operator "$\stackrel{def}{=}$" can be used to associate names with behaviour. Its usage to define single components has been shown in the above examples. Moreover, it can be used for *system definition* which specifies how the system is constructed from the defined components, i.e., how the components cooperate with each other so that they express the behaviour of the system. For example, the printing system with two queues and one printer can be given a name *System*, which is associated with the cooperation between the components *Queue* and *Printer*. That is:

$$System \stackrel{def}{=} (Queue_0 \| Queue_0) \underset{\{print\}}{\bowtie} Printer$$

where $Queue_0$ and *Printer* define the initial behaviour of the corresponding components.

The system states of a PEPA model are the *feasible* combinations of the state of each component of that model. The above printing system example with two queues and one printer has 32 system states if N is 3. For example, the state $(Queue_2 \| Queue_1) \underset{\{print\}}{\bowtie} Printer'$ is one of them.

5.2 Deriving Performance Measures

For any PEPA model, an underlying stochastic process can be generated: a state is associated with a component, and the transitions between the states are defined by the activities between them. Since the duration of a transition in PEPA is exponentially distributed, it has been established that the stochastic process underlying a PEPA model is a CTMC. Performance measures are usually derived from the equilibrium probability vector of the CTMC using the Markov reward structure (MRS) [20].

The equilibrium probabilities of a CTMC can be regarded as the probability that the system is in a state when observed at a random time, or alternatively, can be regarded as the portion of the time the system spends in that state. In MRSs, rewards are associated with states of a Markov process or with transitions between

states, and performance measures are then calculated as the total reward based on the equilibrium system state distribution. If ρ_i is the reward associated with system state s_i, and $\pi(s_i)$ is the equilibrium probability of s_i, then the total reward R is:

$$R = \sum_{s_i \in S} \rho_i * \pi(s_i), \quad (1)$$

where S is the set of all the feasible system states of the PEPA model and ρ_i can be any meaningful value. In this way, various types of measures can be derived from the PEPA model level rather than at the Markov process level.

6 PEPA Models of Network Selection Strategies

In this section, PEPA models that describe the behaviour of a mobile node using different NSSs are presented. Each strategy has its own corresponding PEPA model, and each model consists of three PEPA components that capture the session, handover and network selection behaviour of the mobile node. The same traffic and mobility patterns of the mobile node are used in each model in order that the results depend only on the characteristics of different NSSs. In the following subsections, the PEPA components for the session and handover behaviour of the mobile node are presented first, followed by the PEPA components corresponding to each NSS.

6.1 PEPA Component for Traffic Model

A mobile node's session behaviour is modelled by the component *SESS*. According to Section 3, it can be in an idle or an engaged state.

6.1.1 Idle State of Component *SESS*

In state $SESS_{Idle}$, the component *SESS* can carry out two sets of new session request activities, depending on the position of the mobile node in the 3G-WLAN interworking cell:

- When the mobile node is in the 3G area, new session requests are generated at the rate of λ and all of them are submitted to the 3G RAN (3GRAN). The new session request activities are classified by the session type (activities $sess\_req_{NRT}$ and $sess\_req_{RT}$) and they are generated with the probabilities P_{NRT} and P_{RT} respectively.
- When the mobile node is in the 3G-WLAN area, the new session request activities are further classified by the RAN the requests are submitted to. For example, the activities $sess\_req^{C}_{NRT}$ and $sess\_req^{W}_{RT}$ mean that an NRT session request is submitted to the 3GRAN and an RT session request is submitted to the WLAN RAN (WRAN) respectively. The RAN to which the request is submitted is determined by the PEPA component for the NSS, which will be discussed in Section 6.3.
- Once the new session requests are admitted, the component *SESS* goes to one of the engaged states. The probability of admitting the new session requests is reflected in the cooperation between the component *SESS* and the PEPA component for mobility, which will be discussed in Section 6.2.[1] A new session request may be rejected, and in this case the component *SESS* remains in the idle state. Since the activities corresponding to new session blocking are self-transitions and have no effect on the equilibrium probability of the system states, they are omitted and not defined in the idle state.

The idle state of the component *SESS* is defined as:

$$\begin{aligned} SESS_{Idle} \stackrel{def}{=} & (sess\_req_{NRT},\ P_{NRT} * \lambda).SESS_{NRT} \\ & + (sess\_req_{RT},\ P_{RT} * \lambda).SESS_{RT} \\ & + (sess\_req^{C}_{NRT},\ P_{NRT} * \lambda).SESS_{NRT} \\ & + (sess\_req^{W}_{NRT},\ P_{NRT} * \lambda).SESS_{NRT} \\ & + (sess\_req^{C}_{RT},\ P_{RT} * \lambda).SESS_{RT} \\ & + (sess\_req^{W}_{RT},\ P_{RT} * \lambda).SESS_{RT} \end{aligned}$$

6.1.2 Engaged States of Component *SESS*

Depending on the type of a session, the component *SESS* can be either in the state $SESS_{NRT}$ or in the state $SESS_{RT}$. When the session is completed (activities $sess_{NRT}$ and $sess_{RT}$), or is dropped during a handover (activities *HHO_blk* and *VHO_blk*), the component *SESS* goes back to the idle state.

The engaged states of the component *SESS* are defined as:

$$\begin{aligned} SESS_{NRT} \stackrel{def}{=} & (sess_{NRT},\ \mu_{NRT}).SESS_{Idle} \\ & + (VHO\_blk,\ \top).SESS_{Idle} \\ & + (HHO\_blk,\ \top).SESS_{Idle} \\ SESS_{RT} \stackrel{def}{=} & (sess_{RT},\ \mu_{RT}).SESS_{Idle} \\ & + (VHO\_blk,\ \top).SESS_{Idle} \\ & + (HHO\_blk,\ \top).SESS_{Idle} \end{aligned}$$

6.2 PEPA Component for Mobility Model

A mobile node's handover behaviour is captured by the component *MN*. Moreover, the component *MN* also expresses the session and network selection behaviour of the mobile node, but they are *controlled* by the corresponding PEPA components, as reflected by the symbol "$\top$". Like the component *SESS*, the component *MN* can be in an idle or an engaged state.

6.2.1 Idle States of Component *MN*

The idle states of the component *MN* imply that the mobile node is not communicating and they are characterised by the position of the mobile node in the 3G-WLAN interworking cell. The subscripts $2i-1$ and $2i$ ($i = 1, 2, \ldots, N/2$) are used to denote which phase of the mobility model the mobile node is currently in. The idle states of the component *MN* cooperate with the

1. The activity rates of successful new session requests are in fact not $P_{NRT} * \lambda$ and $P_{RT} * \lambda$ and they are determined by the cooperation.

component *SESS* on different sets of new session request activities according to the mobile node's position:

- In state MN^{Idle}_{2i-1}, the mobile node is in the 3G area. It submits new session requests to the 3GRAN (activities $sess\_req_{NRT}$ and $sess\_req_{RT}$). The probability of admitting a new session request in the 3GRAN is reflected in the parameter P^C_{NA}. For example, by the cooperation between MN^{Idle}_{2i-1} and $SESS_{Idle}$, the rate of the activity $sess\_req_{NRT}$ is actually $P^C_{NA} * P_{NRT} * \lambda$, rather than $P_{NRT} * \lambda$.
- In state MN^{Idle}_{2i}, the mobile node is in the 3G-WLAN area. It submits new session requests to different RANs according to its selection strategy (e.g. activities $sess\_req^C_{NRT}$ and $sess\_req^W_{RT}$). The probability of admitting a new session request to the WRAN is P^W_{NA}. The probabilities of selecting 3GRAN and WRAN are P_C and P_W respectively and they are defined in the PEPA component for NSS. For example, the rate of the activity $sess\_req^W_{RT}$ is actually $P^W_{NA} * P_W * P_{RT} * \lambda$.
- Once new session requests are admitted, the component *MN* goes into an engaged state. As for the component *SESS*, the activities corresponding to new session blocking are omitted since they are self-transitions.
- The mobile node can stay in the idle state and just move within and across the 3G-WLAN interworking cells. The activity $move_{m,n}$ represents the movement of the mobile node from phase m to phase n of its mobility model.

The idle states of the component *MN* are defined as ($i = 1, 2, \ldots, N/2$):

$$
\begin{aligned}
MN^{Idle}_{2i-1} \stackrel{def}{=}\ & (sess\_req_{NRT},\ P^C_{NA} * \top).MN^C_{2i-1} \\
+\ & (sess\_req_{RT},\ P^C_{NA} * \top).MN^C_{2i-1} \\
+\ & (move_{2i-1,2i},\ a_{2i-1} * \upsilon_{2i-1}).MN^{Idle}_{2i} \\
+\ & (move_{2i-1,1},\ b_{2i-1} * \upsilon_{2i-1}).MN^{Idle}_{1} \\
MN^{Idle}_{2i} \stackrel{def}{=}\ & (sess\_req^C_{NRT},\ P^C_{NA} * \top).MN^C_{2i} \\
+\ & (sess\_req^C_{RT},\ P^C_{NA} * \top).MN^C_{2i} \\
+\ & (sess\_req^W_{NRT},\ P^W_{NA} * \top).MN^W_{2i} \\
+\ & (sess\_req^W_{RT},\ P^W_{NA} * \top).MN^W_{2i} \\
+\ & (move_{2i,2i+1},\ a_{2i} * \upsilon_{2i}).MN^{Idle}_{2i+1} \\
+\ & (move_{2i,1},\ b_{2i} * \upsilon_{2i}).MN^{Idle}_{1} \\
MN^{Idle}_{N} \stackrel{def}{=}\ & (sess\_req^C_{NRT},\ P^C_{NA} * \top).MN^C_{N} \\
+\ & (sess\_req^C_{RT},\ P^C_{NA} * \top).MN^C_{N} \\
+\ & (sess\_req^W_{NRT},\ P^W_{NA} * \top).MN^W_{N} \\
+\ & (sess\_req^W_{RT},\ P^W_{NA} * \top).MN^W_{N} \\
+\ & (move_{N,1},\ b_N * \upsilon_N).MN^{Idle}_{1}
\end{aligned}
$$

6.2.2 Engaged States of Component MN

The engaged states of the component *MN* imply that the mobile node is communicating and they are characterised by both the position of the mobile node in the 3G-WLAN interworking cell and the RAN it is currently connected to. The superscripts C and W are used to denote the mobile node is using the 3GRAN and WRAN respectively. The engaged states of the component *MN* cooperate with the component *SESS* on the session holding activities and with the component for the NSS on the network selection activities.

- In state MN^C_{2i-1}, the mobile node is in the 3G area and is connected to the 3GRAN. When it moves into another 3G-WLAN interworking cell (i.e. from phase $2i-1$ to phase 1), it performs a horizontal handover (HHO) (activity $HHO_{2i-1,1}$). When it moves into the 3G-WLAN area of the same cell (i.e. from phase $2i-1$ to phase $2i$), it either performs a VHO to the WRAN (activity $VHO_{2i-1,2i}$), or performs no handover (NHO) and keeps its connection to the 3GRAN (activity $NHO_{2i-1,2i}$). Which type of handover is performed is decided by its NSS and is controlled by the PEPA component for NSS.[2]
- In state MN^C_{2i}, the mobile node is in the 3G-WLAN area and is connected to the 3GRAN. When it moves into another 3G-WLAN interworking cell (i.e. from phase $2i$ to phase 1), it performs a HHO (activity $HHO_{2i,1}$). When it moves into the 3G area of the same 3G-WLAN interworking cell (i.e. from phase $2i$ to phase $2i+1$), no handover is required (activity $NHO_{2i,2i+1}$).
- In state MN^W_{2i}, the mobile node is in the 3G-WLAN area and is connected to the WRAN. It always performs a VHO (activities $VHO_{2i,1}$ and $VHO_{2i,2i+1}$) when it moves out of the 3G-WLAN area (i.e. from phase $2i-1$ to phase 1 and from phase $2i-1$ to phase $2i$). To help understand the above different types of handovers, Fig. 3 illustrates the handover related transitions between the engaged states.

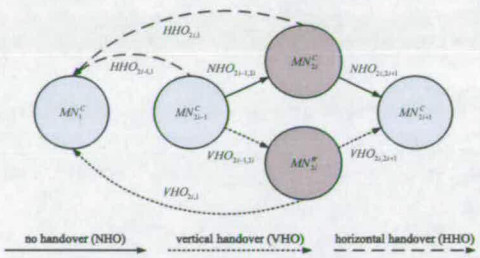


Fig. 3. Different types of handovers between the engaged states

- When the mobile node finishes its session (activities $sess_{NRT}$ and $sess_{RT}$), the component *MN* returns to an idle state. Generally, the NRT sessions are aware of the different data rates provided by different RATs. Therefore, the factors R^C_{NRT} and R^W_{NRT} are used to adjust the duration of the NRT sessions when the mobile node is connected to the 3GRAN and WRAN respectively. For example, the duration of the NRT session in the 3GRAN is $1/(R^C_{NRT} * \mu_{NRT})$.

2. Note that since the network selection only happens when the mobile node moves from the 3G area into the 3G-WLAN area, only $NHO_{2i-1,2i}$, $VHO_{2i-1,2i}$ and *VHO_blk* are *network selection related* activities during a handover.

- When the horizontal and vertical handover requests of the mobile node are blocked (activities *HHO_blk* and *VHO_blk*), the component *MN* also goes back to an idle state. The probabilities of admitting and blocking handover sessions in the 3GRAN and WRAN are P^C_{HA}, P^C_{HB}, P^W_{HA}, and P^W_{HB} respectively.

The engaged states of the component *MN* are defined as ($i = 1, 2, \ldots, N/2$):

$$
\begin{aligned}
MN^C_{2i-1} \stackrel{def}{=}\ & (sess_{NRT},\ R^C_{NRT} * \top).MN^{Idle}_{2i-1} \\
+\ & (sess_{RT},\ \top).MN^{Idle}_{2i-1} \\
+\ & (HHO_{2i-1,1},\ P^C_{HA} * b_{2i-1} * \upsilon_{2i-1}).MN^C_1 \\
+\ & (HHO\_blk,\ P^C_{HB} * b_{2i-1} * \upsilon_{2i-1}).MN^{Idle}_1 \\
+\ & (NHO_{2i-1,2i},\ a_{2i-1} * \upsilon_{2i-1}).MN^C_{2i} \\
+\ & (VHO_{2i-1,2i},\ P^W_{HA} * a_{2i-1} * \upsilon_{2i-1}).MN^W_{2i} \\
+\ & (VHO\_blk,\ P^W_{HB} * a_{2i-1} * \upsilon_{2i-1}).MN^{Idle}_{2i} \\
MN^C_{2i} \stackrel{def}{=}\ & (sess_{NRT},\ R^C_{NRT} * \top).MN^{Idle}_{2i} \\
+\ & (sess_{RT},\ \top).MN^{Idle}_{2i} \\
+\ & (HHO_{2i,1},\ P^C_{HA} * b_{2i} * \upsilon_{2i}).MN^C_1 \\
+\ & (HHO\_blk,\ P^C_{HB} * b_{2i} * \upsilon_{2i}).MN^{Idle}_1 \\
+\ & (NHO_{2i,2i+1},\ a_{2i} * \upsilon_{2i}).MN^C_{2i+1} \\
MN^W_{2i} \stackrel{def}{=}\ & (sess_{NRT},\ R^W_{NRT} * \top).MN^{Idle}_{2i} \\
+\ & (sess_{RT},\ \top).MN^{Idle}_{2i} \\
+\ & (VHO_{2i,1},\ P^C_{HA} * b_{2i} * \upsilon_{2i}).MN^C_1 \\
+\ & (VHO\_blk,\ P^C_{HB} * b_{2i} * \upsilon_{2i}).MN^{Idle}_1 \\
+\ & (VHO_{2i,2i+1},\ P^C_{HA} * a_{2i} * \upsilon_{2i}).MN^C_{2i+1} \\
+\ & (VHO\_blk,\ P^C_{HB} * a_{2i} * \upsilon_{2i}).MN^{Idle}_{2i+1} \\
MN^C_N \stackrel{def}{=}\ & (sess_{NRT},\ R^C_{NRT} * \top).MN^{Idle}_N \\
+\ & (sess_{RT},\ \top).MN^{Idle}_N \\
+\ & (HHO_{N,1},\ P^C_{HA} * \upsilon_N).MN^C_1 \\
+\ & (HHO\_blk,\ P^C_{HB} * \upsilon_N).MN^{Idle}_1 \\
MN^W_N \stackrel{def}{=}\ & (sess_{NRT},\ R^W_{NRT} * \top).MN^{Idle}_N \\
+\ & (sess_{RT},\ \top).MN^{Idle}_N \\
+\ & (VHO_{N,1},\ P^C_{HA} * \upsilon_N).MN^C_1 \\
+\ & (VHO\_blk,\ P^C_{HB} * \upsilon_N).MN^{Idle}_1
\end{aligned}
$$

6.3 PEPA Component for Each Network Selection Strategy

A mobile node's network selection behaviour is controlled by the component *NS*, and a different component *NS* is required for each strategy. The component *NS* is designed to synchronise with the component *SESS* and the component *MN* on the *network selection related* activities. In this way, the mobile node's choice of a certain RAN can easily be controlled by enabling and disabling the synchronisation of the corresponding activities.

Different NSSs are classified into two groups: non-deterministic and deterministic. Non-deterministic strategies choose the RAN according to some on-line measures. On the other hand, deterministic strategies choose the RAN according to a predefined procedure. In this work, PEPA models of three types of selection strategies are built, namely general, WLAN-first and service-based. The general NSS model is for the common non-deterministic strategies and it embeds randomness in network selection. The WLAN-first and service-based NSSs models are for the specific deterministic strategies as their names suggest. The component *NS* and the system definition for each type of strategy are described in the following subsections.

6.3.1 General Network Selection Strategy

Using the general strategy, the mobile node chooses the 3GRAN and WRAN with non-zero probabilities P_C and P_W respectively. One example is the random strategy which chooses the 3GRAN and the WRAN with equal probabilities. Another example is the strategy that is based on the relative received signal strength (RRSS) [10]. No matter how the non-deterministic strategies are designed, the PEPA component for the general network selection NS^G should enable selecting both RANs for new and handover sessions. The component NS^G is defined as:

$$
\begin{aligned}
NS^G \stackrel{def}{=}\ & (sess\_req^C_{NRT},\ P_C * \top).NS^G \\
+\ & (sess\_req^C_{RT},\ P_C * \top).NS^G \\
+\ & (sess\_req^W_{NRT},\ P_W * \top).NS^G \\
+\ & (sess\_req^W_{RT},\ P_W * \top).NS^G \\
+\ & (NHO_{2i-1,2i},\ P_C * \top).NS^G \\
+\ & (VHO_{2i-1,2i},\ P_W * \top).NS^G \\
+\ & (VHO\_blk,\ P_W * \top).NS^G
\end{aligned}
$$

6.3.2 WLAN-first Network Selection Strategy

Using the WLAN-first strategy, the mobile node always chooses the WRAN whenever it is available. WRAN is usually preferred because of its high bandwidth, small delay and low cost. This strategy can be implemented by disabling the activities corresponding to selecting the 3GRAN for both new and handover sessions. The PEPA component for WLAN-first network selection NS^{WF} is defined as:

$$
\begin{aligned}
NS^{WF} \stackrel{def}{=}\ & (sess\_req^W_{NRT},\ P_W * \top).NS^{WF} \\
+\ & (sess\_req^W_{RT},\ P_W * \top).NS^{WF} \\
+\ & (VHO_{2i-1,2i},\ P_W * \top).NS^{WF} \\
+\ & (VHO\_blk,\ P_W * \top).NS^{WF}
\end{aligned}
$$

6.3.3 Service-based Network Selection Strategy

Using the service-based strategy, the mobile node chooses the RAN according to the type of its ongoing session. For example, the mobile node may choose the WRAN for NRT sessions and 3GRAN for RT sessions, because the NRT sessions can take advantage of the higher data rate provided by WRAN and the RT sessions will experience less handovers when choosing 3GRAN.

Since the mobile node makes its decision based on the session type, it would be better to let the component *SESS* control the network selection behaviour. For that reason, the engaged states of the component *SESS* are modified to implement service-based network selection during *handover*. For the NRT sessions, since the mobile node always performs a VHO from the 3GRAN to the WRAN, the activity $VHO_{2i-1,2i}$ is enabled. On the other hand, since the RT sessions always use the 3GRAN and no vertical handover is required, the activity $NHO_{2i-1,2i}$ is enabled and the activity *VHO_blk* is not needed. The component *SESS* for the service-based strategy is modified and renamed as $SESS^{SB}$, whose idle state is

the same as that of $SESS$ whereas its engaged states are defined as:

$$\begin{aligned} SESS^{SB}_{NRT} &\stackrel{def}{=} (sess_{NRT},\ \mu_{NRT}).SESS^{SB}_{Idle} \\ &+ (VHO_{2i-1,2i},\ P_W * \top).SESS^{SB}_{NRT} \\ &+ (VHO\_blk,\ P_W * \top).SESS^{SB}_{Idle} \\ &+ (HHO\_blk,\ \top).SESS^{SB}_{Idle} \\ SESS^{SB}_{RT} &\stackrel{def}{=} (sess_{RT},\ \mu_{RT}).SESS^{SB}_{Idle} \\ &+ (NHO_{2i-1,2i},\ P_C * \top).SESS^{SB}_{RT} \\ &+ (HHO\_blk,\ \top).SESS^{SB}_{Idle} \end{aligned}$$

The network selection for a new session request is still implemented in the component NS^{SB} by enabling corresponding activities.[3] The component NS^{SB} is defined as:

$$\begin{aligned} NS^{SB} &\stackrel{def}{=} (sess\_req^{C}_{RT},\ P_C * \top).NS^{SB} \\ &+ (sess\_req^{W}_{NRT},\ P_W * \top).NS^{SB} \end{aligned}$$

6.4 System Definition for Each Network Selection Strategy

The system definitions of the PEPA models of the above three strategies have the same structure: the components NS^G (NS^{WF}), and $SESS$ cooperate with the component MN so as to control the session and network selection behaviour of the mobile node. The cooperation sets in the system definitions of the general and WLAN-first strategies are the same, while those for the service-based strategy are different from the others. The three PEPA models are defined as:

$$\begin{aligned} NSS^{G} &\stackrel{def}{=} SESS_{Idle} \underset{L_1}{\bowtie} MN_1^{Idle} \underset{L_2}{\bowtie} NS^{G}, \\ NSS^{WF} &\stackrel{def}{=} SESS_{Idle} \underset{L_1}{\bowtie} MN_1^{Idle} \underset{L_2}{\bowtie} NS^{WF}, \\ NSS^{SB} &\stackrel{def}{=} SESS^{SB}_{Idle} \underset{L_3}{\bowtie} MN_1^{Idle} \underset{L_4}{\bowtie} NS^{SB}, \end{aligned}$$

where the cooperation sets are

$$\begin{aligned} L_1 &= \{sess\_req_{NRT}, sess\_req_{RT}, sess\_req^{C}_{NRT}, sess\_req^{W}_{NRT}, \\ &\quad sess\_req^{C}_{RT}, sess\_req^{W}_{RT}, sess_{NRT}, sess_{RT}, HHO\_blk, VHO\_blk\}, \\ L_2 &= \{sess\_req^{C}_{NRT}, sess\_req^{W}_{NRT}, sess\_req^{C}_{RT}, sess\_req^{W}_{RT}, \\ &\quad NHO_{2i-1,2i}, VHO_{2i-1,2i}, VHO\_blk\}, \\ L_3 &= \{sess\_req_{NRT}, sess\_req_{RT}, sess\_req^{C}_{NRT}, sess\_req^{W}_{NRT}, \\ &\quad sess\_req^{C}_{RT}, sess\_req^{W}_{RT}, sess_{NRT}, sess_{RT}, HHO\_blk, VHO\_blk, \\ &\quad NHO_{2i-1,2i}, VHO_{2i-1,2i}\}, \\ L_4 &= \{sess\_req^{C}_{NRT}, sess\_req^{W}_{NRT}, sess\_req^{C}_{RT}, sess\_req^{W}_{RT}\}. \end{aligned}$$

We denote a system state of a PEPA model as $s_k^{A,B}$, where k, A and B represent the mobile node's phase of its mobility model, the RAN it is connected to, and the type of the session it is engaged in respectively. For example, $s_3^{C,RT}$ means the mobile node is in phase 3 and is connected to the 3GRAN for an RT session. Moreover, s_k^C and s_k^W are the unions of system states and they are defined as $s_k^C = s_k^{C,NRT} \cup s_k^{C,RT}$ and $s_k^W = s_k^{W,NRT} \cup s_k^{W,RT}$. Note that the system states of all the models are the *feasible* combinations of the state of each component of that model. For example, the model of the WLAN-first NSS does not have the states $s_{2l}^{C,RT}$ and $s_{2l}^{C,NRT}$, and the model of the service-based NSS does not have the states $s_{2l}^{W,RT}$ and $s_{2l}^{C,NRT}$, where $l = 1, 2, \cdots, N/2$.

3. In fact, the network selection for a new session request can also be implemented by modifying the idle state of the component $SESS$. The component NS^{SB} is deliberately used so that the model has the same structure as the other models.

6.5 Performance Measures

In this work, we investigate three performance measures, namely average throughput, RAN blocking probability and handover rate.

Average Throughput: The average throughput is defined as the mean data rate that can be achieved by a mobile node during its communication. To derive this measure, the first step is to obtain the percentages of time the mobile node spends using different RATs for different types of sessions. Therefore, four types of engaged times can be defined as follows:

$$\begin{aligned} T_{C,NRT} &= \sum_{i=1}^{N} \pi(s_i^{C,NRT}), \quad T_{C,RT} = \sum_{i=1}^{N} \pi(s_i^{C,RT}), \\ T_{W,NRT} &= \sum_{i=1}^{N/2} \pi(s_{2i}^{W,NRT}), \quad T_{W,RT} = \sum_{i=1}^{N/2} \pi(s_{2i}^{W,RT}), \end{aligned} \tag{2}$$

where $\pi(s_k^{A,B})$ is the equilibrium probability of system state $s_k^{A,B}$. Then the total percentage of time the mobile node is in the engaged states is:

$$T_{Engaged} = T_{C,NRT} + T_{C,RT} + T_{W,NRT} + T_{W,RT}. \tag{3}$$

Based on the above definitions, the average throughput is calculated as the weighted sum of the proportions of the different engaged times to the total engaged time, where the weights are the corresponding data rates of different RATs. That is:

$$\begin{aligned} THP =& D^C_{NRT} * \frac{T_{C,NRT}}{T_{Engaged}} + D^C_{RT} * \frac{T_{C,RT}}{T_{Engaged}} \\ &+ D^W_{NRT} * \frac{T_{W,NRT}}{T_{Engaged}} + D^W_{RT} * \frac{T_{W,RT}}{T_{Engaged}}, \end{aligned} \tag{4}$$

where D^C_{NRT}, D^C_{RT}, D^W_{NRT} and D^W_{RT} are the data rates that can be achieved by the mobile node when it uses the 3G RAT (3GRAT) and WLAN RAT (WRAT) for NRT and RT sessions respectively.

RAN Blocking Probability: Although the blocking probabilities of different RANs can be regarded as independent input parameters to the PEPA models, in this work an approach which utilises the PEPA models to derive the blocking probabilities is presented.

Handover Rate: The handover rate is defined as the mean number of handover attempts performed by the mobile node per unit time.

7 Derivation of RAN Blocking Probability and Handover Rate

For a certain type of NSS, the blocking probabilities of the 3GRAN and WRAN are derived from the interaction

between its PEPA model and a resource consumption model corresponding to that type of NSS. In this procedure, the horizontal and vertical handover rates of the mobile node are obtained at the same time. In the following subsections, the mathematical expressions of the RAN blocking probability and handover rate are presented first, followed by an iterative method to derive them.

7.1 RAN Blocking Probability

To derive blocking probabilities of RANs, a two-dimensional continuous-time Markov chain (2D-CTMC) is used to model the resource consumption of a 3G-WLAN interworking cell. The state of the 2D-CTMC is denoted by two nonnegative integers (c, w), where c and w are the numbers of engaged users in the 3GRAN and WRAN respectively. For WLAN cells which overlap with two adjacent cellular cells, their resources are assumed to be shared by both 3G-WLAN interworking cells. That is, the changes in the number of users of these WLAN cells are reflected in the changes of the states of their spanning 3G-WLAN interworking cells. As shown in Fig. 4, there are five types of events that change the state of the 2D-CTMC and they are described as follows:

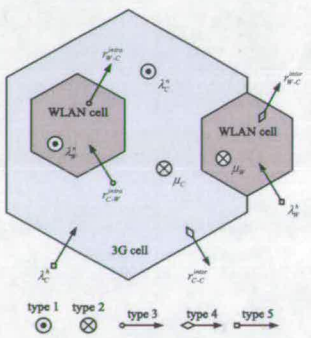


Fig. 4. Five types of events that change the state of the 2D-CTMC

- **type 1**: New sessions requests are generated in the 3GRAN and the WRAN, and their rates are denoted as λ_C^n and λ_W^n respectively.
- **type 2**: Sessions are completed and resources in the 3GRAN and the WRAN are released, and their rates are denoted as μ_C and μ_W respectively.
- **type 3**: Sessions are *internally* handed over *between* the 3GRAN and the WRAN. Their rates are denoted as r_{C-W}^{intra} and r_{W-C}^{intra} respectively.
- **type 4**: Sessions are *externally* handed over *out of* the 3GRAN and the WRAN. Their rates are denoted as r_{C-C}^{inter} and r_{W-C}^{inter} respectively.
- **type 5**: Sessions are *externally* handed over *into* the 3GRAN and WRAN and their rates are denoted as λ_C^h and λ_W^h respectively.

According to the events described above, the state-transition diagram of the 2D-CTMC can be generated. Fig. 5 shows the outward transitions of a non-boundary state (c, w) of the 2D-CTMC. The whole state-transition diagram of the 2D-CTMC can be constructed straightforwardly.

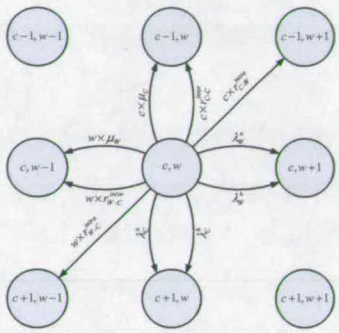


Fig. 5. Outward transitions from a non-marginal state of the 2D-CTMC

For each type of NSS, the rates of the transitions corresponding to the five types of events are calculated as follows:

- **type 1**: Assume that the mobile nodes in the 3G-WLAN interworking cell are uniformly distributed and let A_C and A_W denote the coverage percentage of 3G area and the 3G-WLAN area respectively. For the general and WLAN-first strategies, λ_C^n and λ_W^n are calculated as:

$$\lambda_C^n = A_C * \Lambda^n + P_C * A_W * \Lambda^n, \quad \lambda_W^n = P_W * A_W * \Lambda^n, \quad (5)$$

and for the service-based strategy they are calculated as:

$$\lambda_C^n = A_C * \Lambda^n + P_{RT} * A_W * \Lambda^n, \quad \lambda_W^n = P_{NRT} * A_W * \Lambda^n, \quad (6)$$

where Λ^n is the arrival rate of the new session requests of the whole 3G-WLAN interwork cell.

- **type 2**: For the general and WLAN-first strategies, the resources of the 3GRAN and the WRAN can be used by both NRT and RT users. Therefore in both RANs the probabilities that a session is NRT or RT are P_{NRT} or P_{RT} respectively, and the resources holding time in the 3GRAN $(1/\mu_C)$ and the WRAN $(1/\mu_W)$ are calculated as:

$$\frac{1}{\mu_C} = \frac{P_{NRT}}{R_{NRT}^C * \mu_{NRT}} + \frac{P_{RT}}{\mu_{RT}}, \quad \frac{1}{\mu_W} = \frac{P_{NRT}}{R_{NRT}^W * \mu_{NRT}} + \frac{P_{RT}}{\mu_{RT}}. \quad (7)$$

For the service-based strategy, resource consumption is a bit more complex. The resources of the WRAN can only be used by NRT users in the 3G-WLAN area, whereas the resources of the 3GRAN can be used by RT users in the 3G-WLAN area and all the users in the 3G area. To simplify the analysis, it is assumed that the probabilities that a session in the 3GRAN is NRT and RT are still P_{NRT} and P_{RT} respectively. Accordingly, $1/\mu_C$

and $1/\mu_W$ are calculated as:

$$\frac{1}{\mu_C} \approx \frac{P_{NRT}}{R^C_{NRT} * \mu_{NRT}} + \frac{P_{RT}}{\mu_{RT}}, \quad \frac{1}{\mu_W} = \frac{1}{R^W_{NRT} * \mu_{NRT}}. \tag{8}$$

- **type 3 and 4**: r^{intra}_{C-W}, r^{intra}_{W-C}, r^{inter}_{C-C} and r^{inter}_{W-C} are derived by the approach which will be discussed in Section 7.2.
- **type 5**: λ^h_C and λ^h_W are regarded as input parameters.

Once the transition rates are obtained, the generator matrix of the 2D-CTMC can be generated and its equilibrium probability vector can be derived. Then the blocking probabilities of the 3GRAN and the WRAN are calculated as:

$$P^C_B = \sum_{\substack{c=N_C \\ 0 \leqslant w \leqslant N_W}} p(c,w), \quad P^W_B = \sum_{\substack{w=N_W \\ 0 \leqslant c \leqslant N_C}} p(c,w), \tag{9}$$

where $p(c,w)$ is the equilibrium probability of the state (c,w) of the 2D-CTMC, and N_C and N_W are the maximum number of 3G and WLAN users that can be supported in a 3G-WLAN interworking cell respectively.[4] Note that a different 2D-CTMC is required for each NSS since their generator matrices are different. Eq. (9) are the general expressions for all types of selection strategies.

7.2 Handover Rate

Since the handover rate is defined as the mean number of handover *attempts* performed by the mobile node per unit time, it is actually the throughput of the handover activity. Therefore, the handover rate can be calculated using MRS and Eq. (1), where the rewards are equal to the *activity rate* of that type of handover and they are associated with the system states in which that type of handover is enabled. For example, the activity rate of the internal vertical handover from WRAN to 3GRAN is $a_{2i} * v_{2i}$, and the mobile node performs this type of handover in the system states s^W_{2i}.

Four types of handovers are expressed by the mobility model and their corresponding handover rates are calculated as follows:

$$\begin{aligned}
r^{inter}_{C-C} &= \sum_{i=1}^{N/2} \left(b_{2i-1} * v_{2i-1} * \pi(s^C_{2i-1}) + b_{2i} * v_{2i} * \pi(s^C_{2i}) \right), \\
r^{inter}_{W-C} &= \sum_{i=1}^{N/2} b_{2i} * v_{2i} * \pi(s^W_{2i}), \\
r^{intra}_{C-W} &= \sum_{i=1}^{N/2} P_W * a_{2i-1} * v_{2i-1} * \pi(s^C_{2i-1}), \\
r^{intra}_{W-C} &= \sum_{i=1}^{N/2} a_{2i} * v_{2i} * \pi(s^W_{2i}),
\end{aligned} \tag{10}$$

where $\pi(s^C_{2i-1})$, $\pi(s^C_{2i})$ and $\pi(s^W_{2i})$ are the equilibrium probabilities of the corresponding system states of the PEPA models.

4. The capacity of the 3G-WLAN interworking cell is a complex topic and depends on the user data rate requirements, signal-to-noise ratio, etc [21]. The static capacity is used in order to simply the analysis.

7.3 An Iterative Method to Derive RAN Blocking Probability and Handover Rate

As discussed above, the derivation of RAN blocking probabilities $\boldsymbol{P}_B = [P^C_B, P^W_B]$ requires the handover rates $\boldsymbol{R}_H = [r^{intra}_{C-W}, r^{intra}_{W-C}, r^{inter}_{C-C}, r^{inter}_{W-C}]$ be derived from the PEPA models to calculate $\boldsymbol{p}(c,w)$ of the 2D-CTMC model. On the other hand, $\boldsymbol{R}_H$ are calculated from $\pi(s^{A,B}_k)$ of the PEPA models, which requires $\boldsymbol{P}_B$ be derived from the 2D-CTMC model. Therefore, for each type of NSS, its corresponding 2D-CTMC and PEPA models form a closed loop by exchanging $\boldsymbol{P}_B$ and $\boldsymbol{R}_H$, i.e.,

$$\begin{aligned}
F_{CTMC}(\boldsymbol{R}_H): \quad \boldsymbol{R}_H &\xrightarrow{f_{CTMC}} \boldsymbol{p}(c,w) \xrightarrow{(9)} \boldsymbol{P}_B, \\
F_{PEPA}(\boldsymbol{P}_B): \quad \boldsymbol{P}_B &\xrightarrow{f_{PEPA}} \pi(s^{A,B}_k) \xrightarrow{(10)} \boldsymbol{R}_H,
\end{aligned}$$

where f_{CTMC} and f_{PEPA} denote the mathematical computation to obtain the equilibrium probability vectors of the 2D-CTMC and the PEPA models respectively. The two procedures to derive $\boldsymbol{P}_B$ and $\boldsymbol{R}_H$ are denoted as $F_{CTMC}(\boldsymbol{R}_H)$ and $F_{PEPA}(\boldsymbol{P}_B)$ respectively. To solve this implicit problem, an iterative method is designed and it is described in Algorithm 1.

Algorithm 1: Derivation of $\boldsymbol{P}_B$ and $\boldsymbol{R}_H$

Input: all the required parameters
Output: $\boldsymbol{P}_B = [P^C_B, P^W_B]$,
$\boldsymbol{R}_H = [r^{intra}_{C-W}, r^{intra}_{W-C}, r^{inter}_{C-C}, r^{inter}_{W-C}]$.

```
conv = 0 ;                          /* convergence tag */
ε = 10^-30 ;                        /* convergence criteria */
i = 1 ;                             /* iteration counter */
iter = 100 ;                        /* maximum iteration number */
P_B^(1) = [0,0] ;                   /* initialise P_B^(1) */
R_H^(1) = F_PEPA(P_B^(1)) ;         /* calculate R_H^(1) using P_B^(1) */
while (i ≤ iter) or (conv = 0) do
    P_B^(i+1) = F_CTMC(R_H^(i)) ;   /* update P_B^(i+1) using R_H^(i) */
    err = max{P_B^(i+1) − P_B^(i)} ;   /* calculate difference err */
    if (err > ε) then
        R_H^(i+1) = F_PEPA(P_B^(i+1)) ;   /* calculate R_H^(i+1) using P_B^(i+1) */
        R_H = R_H^(i+1) ;           /* update R_H */
        P_B = P_B^(i+1) ;           /* update P_B */
        i = i + 1;
    else
        conv = 1 ;                  /* set the convergence tag */
        R_H = R_H^(i) ;             /* update R_H */
        P_B = P_B^(i) ;             /* update P_B */
    end
end
return P_B, R_H;
```

The convergence speed of the above iterative method is dependent on the parameter setting but very fast. For example, Table 1 lists the numbers of iterations executed to derive results from each model for 10 increasing session durations. The other parameters are set according to traffic pattern 2, which will be described later.

TABLE 1
Numbers of iterations executed to derive results from each model for 10 increasing session durations listed in TABLE 3

Model	Numbers of iterations
Random	[2, 2, 3, 4, 5, 7, 9, 11, 11, 13]
RRSS	[2, 2, 3, 4, 5, 7, 8, 11, 12, 13]
WLAN-first	[2, 2, 3, 4, 5, 6, 8, 10, 12, 13]
Service-based	[2, 2, 3, 4, 5, 6, 8, 10, 12, 13]

TABLE 3
Activity rates of the PEPA models

Rate	Descriptions	Value (second^{-1})
λ	session arrival rate	180^{-1}
υ_1, υ_3	mobility rate of phase 1 and 3	600^{-1}
υ_2, υ_4	mobility rate of phase 2 and 4	474^{-1}, 1200^{-1}
μ_{RT}	rate of RT sessions	$[60 : 60 : 600]^{-1}$
μ_{NRT}	rate of NRT sessions	$[30 : 30 : 300]^{-1}$
P_{NRT}	NRT session generation probability	0.3, 0.7

8 Performance Evaluation

In this work, four strategies are investigated, namely: random, RRSS, WLAN-first and service-based. The effect of different mobility and traffic patterns of a mobile node on the average throughput, RAN blocking probability and handover rate are investigated.

8.1 Parameter Settings

TABLE 2
Parameter settings of the 2D-CTMC and PEPA models

Branching probabilities			
a_1	0.7	b_1	0.3
a_2	0.5	b_2	0.5
a_3	0.3	b_3	0.7
a_4	0	b_4	1
Data rates and data factors			
D^C_{NRT}	2 Mbps [22]	D^W_{NRT}	6 Mbps [21]
D^C_{RT}	2 Mbps	D^W_{RT}	2 Mbps
R^C_{NRT}	1	R^W_{NRT}	3
Capacities, coverage percentages and session arrival rates			
N_C	50	N_W	30
A_C	0.95	A_W	0.05
λ^h_C	1/20	λ^h_W	1/20
Λ^n	1/10		
Network selection probabilities			
Random	$P_C = 0.5$	$P_W = 0.5$	
RRSS	$P_C = 0.4$	$P_W = 0.6$ [10]	
WLAN-first	$P_C = 0$	$P_W = 1$	
Service-based	$P_C = 1$ (for RT session)		
	$P_W = 1$ (for NRT session)		

Before the evaluation, we need to set up the parameters used in the models. Table 2 lists the settings of the parameters used in the 2D-CTMC and PEPA models. They are divided up into four groups: (a) the phase branching probabilities of the mobility model which is assumed to have four phases; (b) the data rates of the 3GRAT and WRAT for the NRT and RT sessions, and the corrresponding data rate factors; (c) the capacity, coverage percentages of the 3GRAN and the WRAN[5], and the session arrival rates of the 3G-WLAN interworking cell; (d) the network selection probabilities of different strategies.

The activity rates used in the PEPA models of different NSSs are listed in Table 3. The duration of the NRT session $1/\mu_{NRT}$ and that of the RT session $1/\mu_{RT}$ are the control parameters in all the evaluation. The duration of the NRT session is measured at 2 Mbps and thus it reflects the traffic volume of the NRT session. The session arrival interval of a mobile node is set to 180 seconds. The mean sojourn time of the mobile node in the 3G area ($1/\upsilon_1$ and $1/\upsilon_3$) is assumed to be the same and is set to 600 seconds. Two mobility patterns and two traffic patterns are investigated, and they are controlled by the mean sojourn time of the mobile node in the 3G-WLAN area ($1/\upsilon_2$ and $1/\upsilon_4$) and the proportion of the NRT session generated by the mobile node (P_{NRT}).

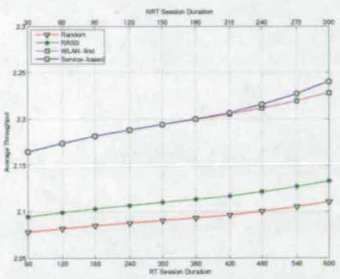
8.2 Effects of Mobility Pattern

The mobility pattern of the mobile node is controlled by its mean sojourn time in the 3G-WLAN area. Two scenarios are considered in the evaluation.

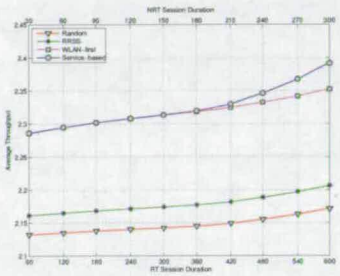
1) In the first scenario, the fluid flow movement model [23] is employed. Based on the formulas developed for the model, and with the assumptions on cell shape and radius, the mean sojourn time of a mobile node in the 3G area and the 3G-WLAN area is 38/30. Then if υ_1 and υ_3 are 600^{-1} then υ_2 and υ_4 are about 474^{-1}.
2) In the second scenario, the mobile node spends a longer time in the 3G-WLAN area and υ_2 and υ_4 are set to 1/1200.

Note that the number of phases, the sojourn times and the branching probabilities of the mobility model used in the evaluation are contrived, more practical values can be estimated from field data using algorithms studied in [16], [18]. NRT and RT sessions are generated at equal probabilities, i.e., $P_{NRT} = P_{RT} = 0.5$. The other

5. We assume there are 5 circular WLAN cells in a circular 3G-WLAN interworking cell, and their radii are set to 100m and 1000m respectively so that the data rates of 6 Mbps and 2 Mbps can be achieved.

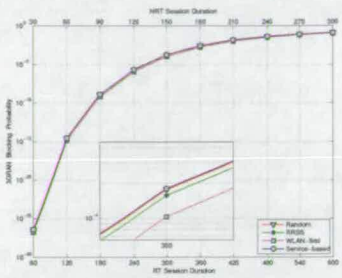


(a) Mobility pattern 1

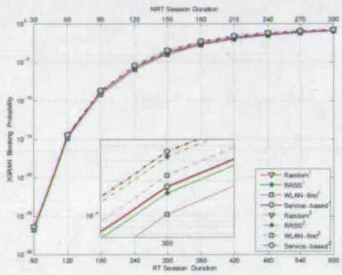


(b) Mobility pattern 2

Fig. 6. The effect of mobility pattern on average throughput



(a) Mobility pattern 1



(b) Mobility pattern 1 and Mobility pattern 2

Fig. 7. The effect of mobility pattern on 3GRAN blocking probability

parameters and activity rates are the same as listed in Table 2 and Table 3. In all the figures, the investigated performance measures are plotted against the session duration. Since there are two types of sessions, two x-axes are used where the top x-axis is the NRT session duration and the bottom x-axis is the RT session duration.

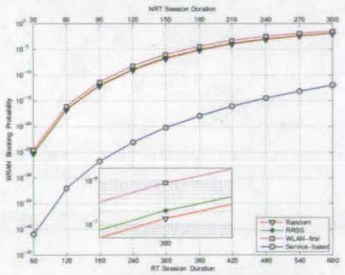
8.2.1 Average Throughput

Fig. 6 shows the average throughput achieved by the mobile node with different mobility patterns. According to Eq. (4), if $D_W^{NRT} = 2$, then the average throughput is always 2 Mbps. Therefore, the average throughput mostly depends on how much time an NRT session uses the WRAT. Fig. 6 indicates that a higher average throughput can be gained by using the strategy with a larger WRAN selection probability, or by simply staying in the 3G-WLAN area for a longer time. Moreover, a longer NRT session duration also results in a higher average throughput, because the mobile node has more opportunity to use the WRAT. An interesting observation is that when NRT session duration is less than 210 seconds, the WLAN-first and service-based strategies have almost the same performance on average throughput but for longer session durations, more of an NRT session in the service-based strategy is spent on the WRAT which results in a clear improvement on the average throughput. This result suggests that the service-based strategy can make the best use of the high data rate of the WRAN, especially for long NRT sessions.

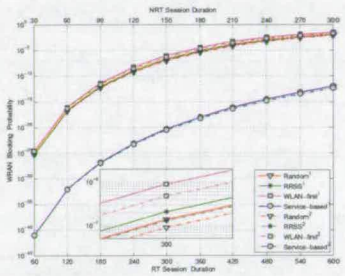
8.2.2 RAN Blocking Probability

Fig. 7 shows the blocking probability of the 3GRAN experienced by the mobile node with different mobility patterns. Note that the y-axis is a logarithmic scale. Since the 2D-CTMC resource consumption model does not include any handover prioritised scheme, the derived blocking probability is for both new and handover session requests. The results indicate that the blocking probability of the 3GRAN mostly depends on its traffic load, which is decided by *how frequently* it is chosen for a session and *how long* the session engages the resources. From Fig. 7(a), it can be observed that the service-based and random strategies are very close and also have the highest blocking probabilities. This is because given $P_{NRT} = P_{RT} = 0.5$, the random and service-based strategies have the same and also the highest probability of using 3GRAN. This can be explained as follows: in the service-based strategy, although only RT sessions choose the 3GRAN, the probability that a session is RT is 0.5; whereas in the random strategy, both types of session can choose the 3GRAN with the probability of 0.5. However, since the approximation made in Eq. (8) is an underestimate[6], the service-based strategy should

6. This is because the RT session, which has a longer duration, should have a proportion larger than P_{RT}.

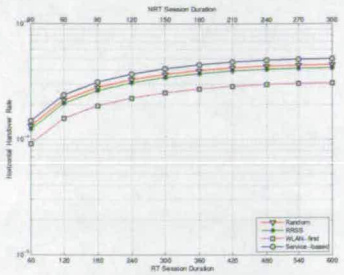


(a) Mobility pattern 1

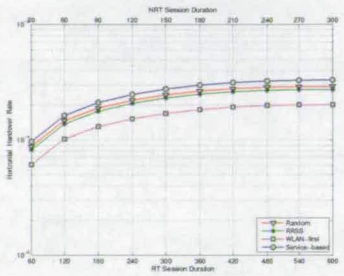


(b) Mobility pattern 1 and Mobility pattern 2

Fig. 8. The effect of mobility pattern on WRAN blocking probability



(a) Mobility pattern 1



(b) Mobility pattern 2

Fig. 9. The effect of mobility pattern on horizontal handover rate

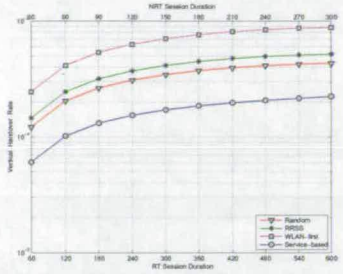
have a clearly higher 3GRAN blocking probability than the random strategy. The RRSS strategy ranks third as all the sessions choose the 3GRAN with the probability of 0.4 and the WLAN-first strategy has the lowest 3GRAN blocking probability. In Fig. 7(b), the effect of the mobile node's mobility is shown. For all of the strategies, a longer stay in the 3G-WLAN area results in a higher 3GRAN blocking probability.

Similarly, the blocking probabilities of the WRAN also depends on its traffic load and the differences between the different strategies are more obvious as shown in Fig. 8. It can be observed that the WRAN blocking probability of the service-based strategy is much smaller than the others. This is because only NRT sessions are allowed to use the WRAT and more importantly they engage the WRAN resources for a shorter time than the RT sessions. Moreover, the mobility of the mobile node has the opposite effect on the WRAN blocking probability to that on the 3GRAN, i.e., a longer stay in the 3G-WLAN area results in a lower WRAN blocking probability.

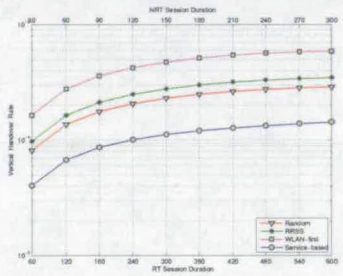
8.2.3 Handover Rate

Fig. 9 shows the horizontal handover rate performed by the mobile node with different mobility patterns. A horizontal handover happens when the engaged mobile node moves across adjacent 3G-WLAN interworking cells. According to Eq. (10), the horizontal handover rate depends on the probability that the mobile node is using the 3GRAT at the time it moves out of its current 3G-WLAN interworking cell. The results indicate that a mobile node using the service-based strategy is the most likely to perform a horizontal handover. This is because the service-based strategy makes the mobile node spend longer in the 3GRAN than the other strategies as explained in Section 8.2.2. The random strategy ranks the second and the WLAN-first strategy has the smallest horizontal handover rate as it uses the 3GRAT less than the other strategies. Given a certain mobility pattern, a longer session duration means the mobile node is more likely to hand over during a session and thus results in a higher handover rate. Moreover, given a certain session duration, mobility can reduce the handover rate for all of the strategies, as shown in Fig. 9(b).

The vertical handover is defined as the handover between different RATs. This rate depends on the probability that the mobile node is using the WRAT at the time it moves out of the 3G-WLAN area, and the probability of choosing the WRAN when it moves into the 3G-WLAN area. Therefore, as shown in Fig. 10, the WLAN-first strategy experiences the most frequent vertical handover whereas the service-based strategy has the lowest vertical handover rate. The effect of the session duration and mobility on the vertical handover rate are the same as those on the horizontal handover rate.

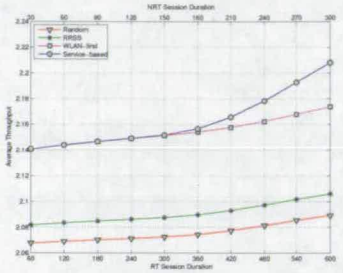


(a) Mobility pattern 1

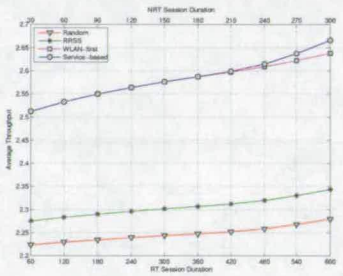


(b) Mobility pattern 2

Fig. 10. The effect of mobility pattern on vertical handover rate



(a) Traffic pattern 1



(b) Traffic pattern 2

Fig. 11. The effect of traffic pattern on average throughput

8.3 Effects of Traffic Pattern

Two traffic scenarios are considered in the evaluation and they are controlled by how frequently the mobile node generates RT and NRT sessions. In the first and second scenario, P_{NRT} is set to 0.3 and 0.7 respectively. All the other parameters are as defined in Table 2 and Table 3. The slow mobility scenario is used, that is, the sojourn time of the mobile node in the 3G-WLAN area is set to 1200 seconds. Again, two x-axes are used in all the figures to show the durations of different types of sessions.

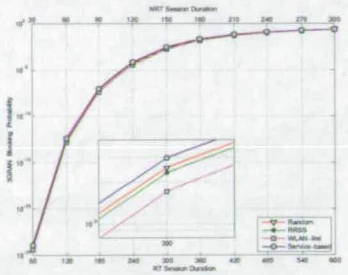
8.3.1 Average Throughput

Similar trends for the average throughput of different strategies as in Section 8.2.1 can be observed in Fig. 11. Moreover, by comparing Fig. 6(b) and Fig. 11, it can be found that a higher NRT probability results in a larger average throughput since in this case there will be more NRT sessions that use the WRAT.

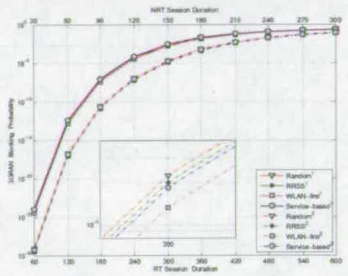
8.3.2 RAN Blocking Probability

Fig. 12 shows the 3GRAN blocking probability experienced by the mobile node with different traffic patterns. In Fig. 12(a) where $P_{NRT} = 0.3$, the service-based strategy has the heaviest traffic load since 70% of the traffic are RT sessions which choose the 3GRAN. The random strategy comes second with all the sessions choosing the 3GRAN with a probability of 0.5, and the WLAN-first strategy has the lowest 3GRAN blocking probability. In the second traffic pattern where $P_{NRT} = 0.7$, the 3GRAN blocking probabilities of all the strategies are reduced as shown in Fig. 12(b), which is mainly because in the parameter settings the duration of an NRT session is shorter than that of an RT session. Therefore, a higher percentage of NRT sessions reduces the resource engagement time of the 3GRAN. Moreover, the 3GRAN blocking probability of the service-based strategy is very sensitive to the traffic pattern and is reduced by a larger amount than the others and is lower than those of the random and RRSS strategies.

The advantage of the service-based strategy on the WRAN blocking probability is very clear as shown in Fig. 13(a) where a larger probability of choosing the WRAN results in a higher WRAN blocking probability. An interesting observation is that at a higher NRT probability, the WRAN blocking probability of the service-based strategy grows whereas those of the other strategies are reduced as shown in Fig. 13(b). The reason is that in the service-based strategy the WRAN resources are only engaged by the NRT sessions whereas in the other strategies the WRAN resources can be engaged by all types of sessions. As a result, a higher NRT probability implies a higher WRAN traffic load in the service-based strategy, whereas in the other strategies this means there will be fewer RT sessions that cannot make use of the high data rate of WRAT and thus engage the WRAN

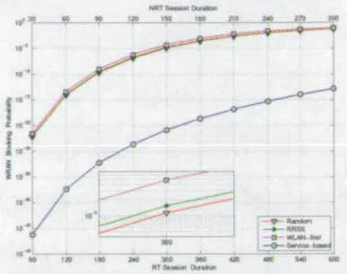


(a) Traffic pattern 1

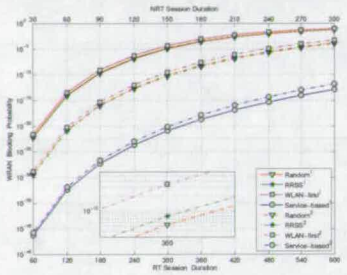


(b) Traffic pattern 1 and Traffic pattern 2

Fig. 12. The effect of traffic pattern on 3GRAN blocking probability



(a) Traffic pattern 1



(b) Traffic pattern 1 and Traffic pattern 2

Fig. 13. The effect of traffic pattern on WRAN blocking probability

resources.

8.3.3 Handover Rate

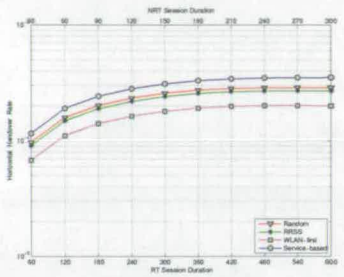
The effect of traffic pattern on the horizontal handover rate is shown in Fig. 14. As can be observed by comparing Fig. 9(b) and Fig. 14, the traffic pattern changes the horizontal handover rate of the random, RRSS and WLAN-first strategies to a small extent since their selection strategies are not based on the type of the session. As explained in Section 8.3.2, a higher percentage of NRT sessions reduces the time the mobile node is connected to the 3GRAN. Therefore, the horizontal handover rates of the these strategies are reduced slightly. On the other hand, since the service-based strategy only allows RT sessions to use the 3GRAT, a lower RT probability means that a mobile node is less likely to be connected to the 3GRAN and thus has a smaller horizontal handover rate. This is why when $P_{NRT} = 0.7$, the horizontal handover rate of the service-based strategy is reduced and is almost the same as that of the random strategy.

A similar effect of the traffic pattern on the vertical handover rate can be observed in Fig. 15. A larger NRT probability results in a higher vertical handover rate in the service-based strategy since the mobile node uses the WRAT more frequently and thus is more likely to be connected to the WRAN. Unlike the horizontal handover rate, the vertical handover rate in the other strategies is more sensitive to the traffic pattern; they decrease as the time the mobile node is connected to the WRAN is reduced at larger NRT probability.

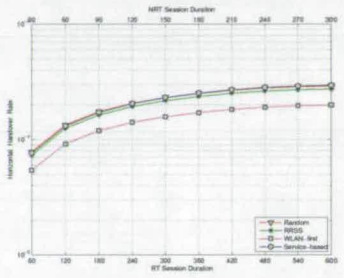
9 Conclusions

To find out the effect of different NSSs on the performance of both mobile nodes and RANs, in this work we investigate a general performance evaluation framework for NSSs. This framework is general because it has an interface to the NSS used by the mobile node and an interface to the resource consumption model of the RANs. Four types of strategies, namely random, RRSS, WLAN-first and service-based, have been evaluated from different perspectives.

The three types of performance measures discussed in this work are meaningful from both the user and the network administrator's perspectives. Both average throughput and handover rate have effect on the quality of service (QoS) perceived by the user. The average throughput is important as it reflects the efficiency of the communication especially for NRT sessions. The handover rate indicates the volume of signalling load and the frequency of service interruption during a session. Therefore a high handover rate should be avoided and in particular vertical handovers, since their cost is higher than that of horizontal handovers due to more involved process. On the other hand, the network administrator may be more concerned about resource utilisation of

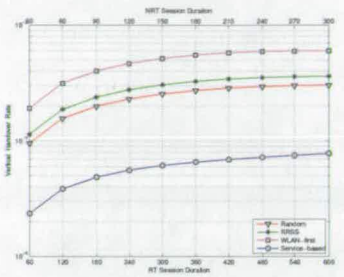


(a) Traffic pattern 1

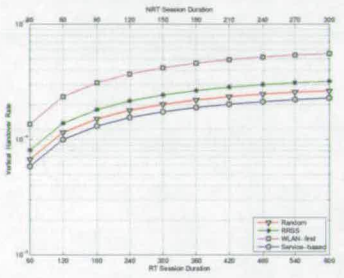


(b) Traffic pattern 2

Fig. 14. The effect of traffic pattern on horizontal handover rate



(a) Traffic pattern 1



(b) Traffic pattern 2

Fig. 15. The effect of traffic pattern on vertical handover rate

RANs and the RAN blocking probability can reflect the traffic loads of different RANs.

The deterministic strategies, such as the WLAN-first and service-based strategies, are easy to implement and a user always knows which RAN is going to be selected. Since the WLAN-first strategy chooses the WLAN whenever it is available, it has the lowest 3GRAN blocking probability and horizontal handover rate, at the expense of having the highest WRAN blocking probability and vertical handover rate. It can also achieve high average throughput but is outperformed by the service-based strategy at long session durations. On the other hand, the service-based strategy makes the best use of the high data rate of the WRAT by only allowing NRT sessions to access the WRAN and consequently has the lowest WRAN blocking probability. Since the service-based strategy is aware of the type of the session, its performance is very sensitive to the traffic pattern of the mobile node. That is, the RT probability is proportional to the 3GRAN blocking probability and horizontal handover rate, and is inversely proportional to the vertical handover rate. As for the non-deterministic strategies, the random and RRSS strategies introduce randomness in network selection and therefore the user will experience uncertainty during the handover. As can be seen from the results, they have more balanced performance on the investigated measures than the deterministic strategies. This phenomenon is likely to extend to other non-deterministic strategies as well since they have intermediate probabilities of choosing the WRAN and 3GRAN.

The effect of the mobility pattern of the mobile node is straightforward. A longer sojourn time in the 3G-WLAN area results in a higher average throughput, and lower horizontal and vertical handover rates. An interesting observation is that the mobile node will experience a higher 3GRAN blocking probability and a lower WRAN blocking probability if the sojourn time in the 3G-WLAN area is longer. As for the effect of the traffic pattern, the attribute of the service-based strategy means it is strongly affected by the traffic pattern, whereas the other strategies are affected simply because the session durations are different when traffic pattern changes.

ACKNOWLEDGEMENTS

The work reported in this paper has formed part of the Ubiquitous Services Core Research Programme of the Virtual Centre of Excellence in Mobile & Personal Communications, Mobile VCE, www.mobilevce.com. This research has been funded by the Industrial Companies who are Members of Mobile VCE, with additional financial support from the UK Governments Technology Strategy Board (previously DTI). Fully detailed technical reports on this research are available to Industrial Members of Mobile VCE.

J. Hillston is also supported by EPSRC Advanced Research Fellowship EP/c543696/01 and EU FET-IST Global Computing 2 project SENSORIA (Software Engineering for Service-Oriented Overlay Computers (IST-3-016004-IP-09)). H. Wang and D. Laurenson acknowledge the support of the Scottish Funding Council for the Joint Research Institute with the Heriot-Watt University which is a part of the Edinburgh Research Partnership.

REFERENCES

[1] J.-C. Cheng and T. Zhang, *IP-based Next-generation Wireless Networks*. John Wiley & Sons, 2004.
[2] "3GPP system to Wireless Local Area Network (WLAN) interworking; System description," 3GPP, Tech. Rep. TS23.234 v7.7.0, Jun. 2008.
[3] M. Stemm and R. H. Katz, "Vertical handoffs in wireless overlay networks," *Springer Mob. Netw. Appl.*, vol. 3, no. 4, pp. 335–350, 1998.
[4] N. Nasser, A. Hasswa, and H. Hassanein, "Handoffs in fourth generation heterogeneous networks," *IEEE Commun. Mag.*, vol. 44, no. 10, pp. 96–103, 2006.
[5] J. McNair and F. Zhu, "Vertical handoffs in fourth-generation multinetwork environments," *IEEE Wireless Commun. Mag.*, vol. 11, no. 3, pp. 8–15, 2004.
[6] E. Stevens-Navarro and V. Wong, "Comparison between vertical handoff decision algorithms for heterogeneous wireless networks," in *Proc. IEEE Vehicular Technology Conference-Spring '06*, 2006, pp. 947–951.
[7] J.-Z. Sun, "A review of vertical handoff algorithms for cross-domain mobility," in *Proc. IEEE International Conference on Wireless Communications and Networking '07*, 2007, pp. 3156–3159.
[8] Q. Song and A. Jamalipour, "Network selection in an integrated wireless LAN and UMTS environment using mathematical modeling and computing techniques," *IEEE Wireless Commun. Mag.*, vol. 12, no. 3, pp. 42–48, 2005.
[9] F. Bari and V. Leung, "Automated network selection in a heterogeneous wireless network environment," *IEEE Netw.*, vol. 21, no. 1, pp. 34–40, 2007.
[10] W. Shen and Q.-A. Zeng, "Cost-function-based network selection strategy in integrated wireless and mobile networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 6, pp. 3778–3788, 2008.
[11] F. Yu and V. Krishnamurthy, "Optimal joint session admission control in integrated WLAN and CDMA cellular networks with vertical handoff," *IEEE Trans. Mobile Comput.*, vol. 6, no. 1, pp. 126–139, 2007.
[12] A. Hasib and A. O. Fapojuwo, "Analysis of common radio resource management scheme for end-to-end QoS support in multiservice heterogeneous wireless networks," *IEEE Trans. Veh. Technol.*, vol. 57, no. 4, pp. 2426–2439, 2008.
[13] X. Gelabert, J. Perez-Romero, O. Sallent, and R. Agusti, "A Markovian approach to radio access technology selection in heterogeneous multiaccess/multiservice wireless networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 10, pp. 1257–1270, 2008.
[14] J. Farber, S. Bodamer, and J. Charzinski, "Statistical evaluation and modelling of Internet dial-up traffic," in *Proc. SPIE Phonics East Conference: Performance and Control of Network Systems III*, 1999, pp. 112–121.
[15] H.-K. Choi and J. O. Limb, "A behavioral model of Web traffic," in *Proc. IEEE International Conference on Network Protocols '99*, 1999, pp. 327–334.
[16] Y. Sasaki, H. Imai, M. Tsunoyama, and I. Ishii, "Approximation of probability distribution functions by coxian distribution to evaluate multimedia systems," *Syst. Comput. Japan*, vol. 35, no. 2, pp. 16–24, 2004.
[17] A. H. Zahran, B. Liang, and A. Saleh, "Mobility modeling and performance evaluation of heterogeneous wireless networks," *IEEE Trans. Mobile Comput.*, vol. 7, no. 8, pp. 1041–1056, 2008.
[18] A. Thummler, P. Buchholz, and M. Telek, "A novel approach for phase-type fitting with the EM algorithm," *IEEE Trans. Dependable Secure Comput.*, vol. 3, no. 3, pp. 245–258, 2006.
[19] J. Hillston, *A Compositional Approach to Performance Modelling*. Cambridge University Press, 1996.
[20] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. John Wiley & Sons, 1998.
[21] J. P. Romero, O. Sallent, R. Agusti, and M. A. Diaz-Guerra, *Radio Resource Management Strategies in UMTS*. John Wiley & Sons, 2005.
[22] J. Kalliokulju, P. Meche, M. J. Rinne, J. Vallstrom, P. Varshney, and S.-G. Haggman, "Radio access selection for multistandard terminals," *IEEE Commun. Mag.*, vol. 39, no. 10, pp. 116–124, 2001.
[23] Q.-A. Zeng and D. P. Agrawal, "Modeling and efficient handling of handoffs in integrated wireless mobile networks," *IEEE Trans. Veh. Technol.*, vol. 51, no. 6, pp. 1469–1478, 2002.

Hao Wang received a B.Eng. degree in Information Engineering from the Southeast University, China, in 2003 and a M.Sc. degree in Signal Processing and Communications from The University of Edinburgh in 2005. He is currently working towards a Ph.D. degree in the School of Engineering at The University of Edinburgh. His research interests include developing and performance modelling of mobility and QoS management schemes in mobile and wireless networks.

David I. Laurenson obtained a B.Eng. degree in Computer Science and Electronics followed by a Ph.D. in Electronics and Electrical Engineering from The University of Edinburgh in 1990 and 1994 respectively. He was appointed as a Lecturer in 1994 with research interests in mobile radio communications, ranging from propagation issues to ad-hoc networking.

Jan Hillston is Professor of Quantitative Modelling in the School of Informatics at The University of Edinburgh and holds an Advanced Research Fellowship from the Engineering and Physical Sciences Research Council. Her principal research interests are in the use of stochastic process algebras to model and analyse computer systems and the development of efficient solution techniques for such models. Prof. Hillston received the B.A. and M.Sc. degrees in Mathematics from the University of York (UK) and Lehigh University (USA), respectively. After a brief period working in industry, she joined the Department of Computer Science at The University of Edinburgh, as a research assistant in 1989. She received the Ph.D. degree in Computer Science from that university in 1994. Her work on the stochastic process algebra PEPA was recognised by the British Computer Society in 2004 who awarded her the first Roger Needham Award.

Wireless Pers Commun
DOI 10.1007/s11277-009-9724-1

A Reservation Optimised Advance Resource Reservation Scheme for Deploying RSVP in Mobile Environments

Hao Wang · David I. Laurenson · Jane Hillston

Abstract One of the major problems of deploying Resource ReSerVation Protocol (RSVP) in mobile environments is called the advance resource reservation (ARR) problem. Conventional solutions to this problem waste too many network resources and increase the new Quality of Service (QoS) session blocking probability. In this paper, we propose a reservation optimised ARR scheme which constrains the number of advance reservation paths in a subnet and only allows the most eligible mobile nodes to make advance reservations. Furthermore, to evaluate the performance of the schemes, we build Markovian models of different ARR schemes using a formal performance modelling formalism named Performance Evaluation Process Algebra (PEPA). Our results indicate that the proposed reservation optimised ARR scheme can effectively balance the active and passive reservation blocking probabilities and achieves a better utilisation of the network resources, especially when the traffic intensity is high.

Keywords Mobile and wireless networks · Resource management · Performance evaluation · Process algebra · RSVP · QoS

1 Introduction

As many real-time services and multimedia applications become popular, providing guaranteed quality of service (QoS) to Internet users is an important issue for the next generation

H. Wang (✉) · D. I. Laurenson
Institute for Digital Communications, Joint Research Institute for Signal & Image Processing, School of Engineering, The University of Edinburgh, Edinburgh, UK
e-mail: H.Wang@ed.ac.uk

D. I. Laurenson
e-mail: Dave.Laurenson@ed.ac.uk

J. Hillston
Laboratory for Foundations of Computer Science, School of Informatics, The University of Edinburgh, Edinburgh, UK
e-mail: Jane.Hillston@ed.ac.uk

Published online: 23 April 2009

of traffic management. One of the proposed solutions is the Integrated Service (IntServ) [1] that utilises a signalling protocol such as Resource ReSerVation Protocol (RSVP) [2] to control end-to-end packet delay. Unlike the top-down provisioning QoS mechanisms such as Differentiated Services (DiffServ) [3] which tends to offer soft-quality guarantees, the signalling based IntServ provides dynamic configuration of the network devices and offers quantifiable end-to-end high-quality guarantees required for multimedia applications [4]. Therefore, although RSVP exhibits the scalability problem [5], it is still one of the most valuable components in a broader set of QoS technologies [6].

RSVP is a receiver-oriented resource reservation setup protocol for simplex data flows. It can be used by a host to request specific amount of resources from the network and by routers to establish and maintain the required QoS. However, RSVP is initially designed for wired networks and performs poorly due to the node mobility in mobile wireless networks [7]. When a mobile node changes its point of attachment to the network, e.g. performs a network layer handover from one subnet to another, it needs to change its IP address. Although the network connectivity of the mobile node can be maintained by the mobility management (MM) protocols such as Mobile IPv6 [8] and Hierarchical Mobile IPv6 [9], the end-to-end QoS is not guaranteed since the reservation path has to be re-established along the new data flow path between the mobile node and its correspondents after handover. Therefore, the integration of mobility and QoS provisioning in mobile wireless networks is a real challenge and a lot of work has been done on deploying RSVP in mobile environments. These proposals are designed to reduce the handover reservation re-establishment delay by localising the RSVP signalling and to avoid the handover reservation re-establishment blocking by making reservations in advance. For a detailed survey of mechanisms combining mobility and QoS managements, refer to [7,10,11] and references therein.

In this paper, we study the side effects of the conventional advance resource reservation (ARR) schemes and propose a reservation optimised ARR scheme which combines resource reservation and call admission control (CAC) mechanisms. The proposed scheme is designed to improve the performance of existing ARR schemes and it can be directly integrated into them. The fundamental purpose of this scheme is to restrict the amount of advance reserved resources in a subnet since allowing too many advance reservation paths is a waste of network resources from the QoS traffic's perspective. Furthermore, to make the advance resource reservation more effective, this scheme takes account of the traffic and mobility patterns of the mobile nodes and only allows the most eligible mobile nodes to reserve resources in advance.

To assess the performance of different ARR schemes, Markovian models of them are built. Constructing Markovian models of complex systems such as these, using traditional performance modelling techniques like queueing networks and stochastic Petri nets is time-consuming, error prone, and not guaranteed to give a mathematically tractable solution. Instead we use a stochastic process algebra, namely Performance Evaluation Process Algebra (PEPA), that is suitable for modelling such complex systems to build the Markovian models. Furthermore, our models are independent of the specific implementations of the ARR schemes and capture the essential characteristics underlying them. From the PEPA models, we derive important performance measures including the blocking probabilities of different types of reservation requests and the mean numbers of different types of reservation paths in a subnet, which demonstrate the advantages of the proposed ARR scheme.

The rest of this paper is structured as follows. In Sect. 2, we present the background of the issues of RSVP in mobile environments and then review the most representative ARR schemes. In Sect. 3, we illustrate the proposed reservation optimised ARR scheme and its operation procedure in detail. In Sect. 4, we give a short introduction to the PEPA formalism

and in Sect. 5 the PEPA models of different ARR schemes are described. We analyse the performance of the different ARR schemes in Sect. 6 and give our conclusions in Sect. 7.

2 Background and Related Work

In this section, we first present the background of the issues of RSVP in mobile environments and then review the most representative work on addressing the issue of advance resource reservation.

2.1 RSVP in Mobile Environments

Resource reservation using conventional RSVP exhibits a lot of deficiencies in mobile environments. This is because in RSVP the resource reservation is identified by the IP addresses of the communicating ends, and a mobile node must re-establish its reservation path after a network layer handover. One problem of this reservation re-establishment is that it causes an interruption to the QoS session due to the reservation setup delay and in the worst case the QoS session has to be terminated if it is very sensitive to packet delay. Therefore, to minimise the effect of node mobility, it is required to localise the reservation signalling so that the reservation re-establishment delay is reduced [12]. In many practical cases, it is highly likely that the old and new data flow paths between the mobile node and its correspondent before and after the handover overlap, and only a small portion of the end-to-end path changes. Therefore, it is beneficial to limit the extent of the reservation re-establishment to the affected part of the end-to-end path. Several approaches of localising reservation re-establishment have been proposed and the details can be found in [13–16].

Another problem of deploying RSVP in mobile environments is that a mobile node may fail to acquire enough network resources after a handover. When the mobile node changes its data flow path after handover, the congestion levels at the routers along the paths also change [11]. If the new path is overcongested, the available bandwidth along the new path may not be sufficient to satisfy the requirements of the QoS session. To solve this problem, it has been suggested that the required resources be reserved in advance in the subnets that a mobile node may visit. Once the resources are guaranteed before handover, the mobile node can continue its QoS session smoothly after it switches its connectivity to another subnet. In the next subsection, we present a more detailed review of the approaches for reserving resources in advance.

2.2 Advance Resource Reservation Schemes

Advance resource reservation schemes aim to reduce the reservation request blocking probability during handover by reserving resources in advance for a mobile node. Previous proposals for ARR schemes can be classified into two types: *agent-based* and *multicast-based*.

In the agent-based schemes [14, 17–19], there are two types of reservations: *active* and *passive*. A mobile node makes an active reservation path from its current subnet and it makes passive reservation paths from neighbouring subnets that it may visit to the correspondent node. An active reservation path is actively used by the mobile node to carry out its communication and the passive reservation paths are just reserved for the mobile node but not used. When the mobile node hands over into a neighbouring subnet, its passive reservation path in the newly visited subnet is switched to the active state and its old active reservation path is changed to the passive state. In every subnet there is an agent (such as Proxy Agent in [17]

and Mobility Agent in [14]) which is in charge of the resource reservation procedure. The mobile node needs to inform the agents of the neighbouring subnets which it may visit of its reservation information and require them to make passive reservation paths for it. In [14], the passive reservation paths are only established when the mobile node is in the overlapped area of two subnets and intends to perform a handover. To reduce redundant passive reservation paths, some approaches are equipped with the position prediction techniques [18,19]. These techniques estimate the most likely neighbouring subnets that the mobile node may visit according to statistical models or historical moving trajectories, and only allow advance resource reservations in the predicted subnets.

In the multicast-based schemes [16,20], RSVP signalling messages and ordinary data packets are delivered to a mobile node using IP multicast routing. As in the agent-based approaches, there is also an agent in every subnet (such as Mobile Proxy in [16] and QoS Agent in [20]). Before the mobile node starts its communication, the mobile node's local agent and the agents in the neighbouring subnets join a multicast group. Since all the traffic from and to the mobile node must go through these agents, a handover of the mobile node can be regarded as leaving and joining the branches of a multicast tree. The local agent makes a *conventional* reservation path between the two communicating ends and the neighbour agents make *predictive* reservation paths on behalf of the mobile node. These two types of reservation states are essentially the same as the *active* and *passive* states in the agent-based schemes. When the mobile node hands over into a neighbouring subnet, the states of its old and new reservation paths are changed accordingly. Since all the data packets are addressed to the multicast address, the mobile node can continue its communication without interruption when it moves out of its current subnet.

With the help of the ARR schemes, the session interruption during handover is reduced. However, an advance reservation path in a subnet is made exclusively for a certain mobile node and is not actively used by its reserver. Allowing too many advance reservation paths in a subnet will increase the blocking probability of the active reservation requests originating from the mobile nodes in that subnet. The approaches that allow traffic with lower QoS requirements to temporarily use the passive reservation paths [14,16,17,20] are not reliable, since the resources borrowed by a QoS session have to be returned when their reservers reclaim them, which results in an interruption to the borrower. On the other hand, only allowing best-effort traffic to use the passive reservation paths is a waste of network resources from the QoS traffic's point of view. Therefore, putting a restriction on the number of the advance reservation paths in a subnet would be beneficial from the perspective of network utilisation. In fact, a combination of ARR and CAC mechanisms should achieve better performance on managing network resources [10,11,21].

3 The Reservation Optimised Advance Resource Reservation Scheme

In this section, we propose a reservation optimised ARR scheme. This scheme aims to achieve a better utilisation of the network resources by balancing the number of active and passive reservation paths in a subnet. The proposed scheme includes two admission mechanisms: passive reservation limited and session-to-mobility ratio (SMR)-based replacement, both of which can be integrated into existing ARR schemes. The work presented in this paper is not a design of a new signalling procedure for an ARR scheme but is an investigation on an efficient way of utilising network resources using an existing signalling mechanism. In the following subsections, we firstly present the two admission control mechanisms that are used to achieve our goal and then introduce the procedure of our scheme.

3.1 Passive Reservation Limited Mechanism

In the conventional ARR schemes, the resources of a subnet are actively reserved by the mobile nodes in that subnet (namely local mobile nodes) and passively reserved by the mobile nodes in the neighbouring subnets (namely foreign mobile nodes). Since the active and passive reservation requests are treated in the same way, there is no restriction on the number of passive reservation paths in a subnet. Moreover, since allowing too many passive reservation paths is a waste of resources from the perspective of QoS traffic, it is better to give higher priority to the active reservation requests because this type of request implies that there are QoS sessions that cannot start without the requested resources.

We borrow the concept of a *channel*, which is widely used in cellular networks, to model the resources of a subnet. To limit the number of passive reservation paths in a subnet, part of the channels are reserved only for active reservation paths. Therefore, the channels of the subnet are partitioned into two groups: *standard channels* and *dedicated channels*. The only difference between a standard channel and a dedicated channel is that the former can be used for both active and passive reservation paths, while the latter can *only* be used for an active reservation path. To guarantee that the channels are allocated correctly, there is an enhanced agent (EA) in each subnet which monitors the network resources and admits different types of reservation requests. The EA assigns the dedicated or standard channels to the active reservation requests, and only assigns the standard channels to the passive reservation requests. In this way, the number of passive reservation paths in the subnet is limited and hence more resources are available for active reservation requests. Moreover, unlike the conventional ARR schemes, the passive reservation limited mechanism does not allow an active reservation path to change to the passive state. That is, when the local mobile node hands over out of the local subnet, it has to release its active reservation path and instead requests a passive reservation path.

More importantly, in order to avoid over-restricting passive reservation requests, the EA first uses the dedicated channels for active reservation paths. The standard channels are only allocated when all the dedicated channels are engaged. Therefore, if the total number of channels in a subnet is T and the number of standard channels is S, then the maximum number of passive reservation paths in the subnet is S and the EA can accept at least $T - S$ active reservation requests.

3.2 SMR-based Replacement Mechanism

Since only the standard channels of a subnet can be used for passive reservation paths, they are scarce resources from the foreign mobile nodes' point of view. Therefore, an efficient admission strategy is necessary to determine which foreign mobile node is eligible to acquire a standard channel. Since the essential objective of an ARR scheme is to improve the handover performance of a mobile node, it would be better to assign a standard channel to the foreign mobile node which is most likely to handover during a session.[1]

In previous work [22,23] the handover frequency of a mobile node is usually defined by the session-to-mobility ratio, which is the ratio of the mobile node's session arrival rate to its mobility rate. This type of ratio is used to optimise the packet routing and network traffic load in HMIPv6. However, a handover is the behaviour of a mobile node during its communication and it has no direct relationship with the session arrival rate. Therefore, in the proposed SMR-based replacement mechanism a modified form of the ratio, which is defined

[1] All the sessions are assumed to have the same QoS class.

as the ratio of a mobile node's session duration to its residence time in a subnet, is used to characterise the handover likelihood of the mobile node during its communication.

In the SMR-based replacement mechanism, every EA is assumed to be able to get the SMR information of the mobile nodes in its administrating subnet.[2] The mechanism works as follows: Assume there is a foreign mobile node which requests a passive reservation path in the local subnet. The foreign EA which is in charge of that foreign mobile node will inform the local EA of the SMR of the requesting foreign mobile node. Then

- If there are free standard channels in the local subnet, the local EA will allocate one to the foreign mobile node.
- If there is no free standard channel in the local subnet, the local EA will compare and find out whether the SMR of the requesting foreign mobile node is larger than the smallest of the SMRs of the foreign mobile nodes that have already been allocated standard channels. If it is, the foreign mobile node with the smallest SMR is replaced by the requesting foreign mobile node, i.e., the standard channel is re-allocated to the requesting mobile node. Otherwise, the passive reservation request is blocked.

Note that the SMR-based replacement mechanism should not be applied to active reservation requests because this would affect the ongoing QoS sessions. On the other hand, the re-allocation of passive reservation paths has no effect on the QoS sessions of the foreign mobile nodes since they are not actively using them.

3.3 Operation Procedure

In the following, we describe the operation procedure of the reservation optimised ARR scheme from the perspective of the EA of the local subnet. Figure 1 shows the channel allocation procedure of the proposed scheme. The local EA receives two types of reservation requests.

- When the local EA receives an active reservation request from a local mobile node, it will allocate a free channel to that local mobile node (a dedicated channel is chosen first, or if one is not available, then a standard channel). When the local mobile node finishes its session, its active reservation path is released. Moreover, it also releases its active reservation path when it hands over out of the local subnet and instead it sends a passive reservation request to the local EA.
- When the local EA receives a passive reservation request from a foreign mobile node, it tries to allocate a standard channel to the foreign mobile node. The allocation procedure is described in Sect. 3.2. If the foreign mobile node fails to obtain a passive reservation path in the local subnet, it has to request an active reservation path when it hands over into the local subnet.

In brief, the reservation optimised ARR scheme is a CAC enhanced solution to the advance resource reservation problem. The CAC is carried out by the EA in each subnet by managing network resources with the considerations of the type of requests and the mobility characteristics of the mobile nodes. The motivation for integrating the CAC algorithm is to restrict the number of passive reservation paths in a subnet and only allow the most eligible mobile nodes to acquire them. In this work, a mobile node is considered to be more eligible if it has a larger SMR, with the assumption that all the sessions are of the same QoS class. However, in a broader sense, different QoS classes should also be considered and it is a very important

[2] This can be achieved by receiving information messages from the mobile nodes, or by employing some statistical or history-based prediction algorithms. These approaches are beyond the scope of this work.

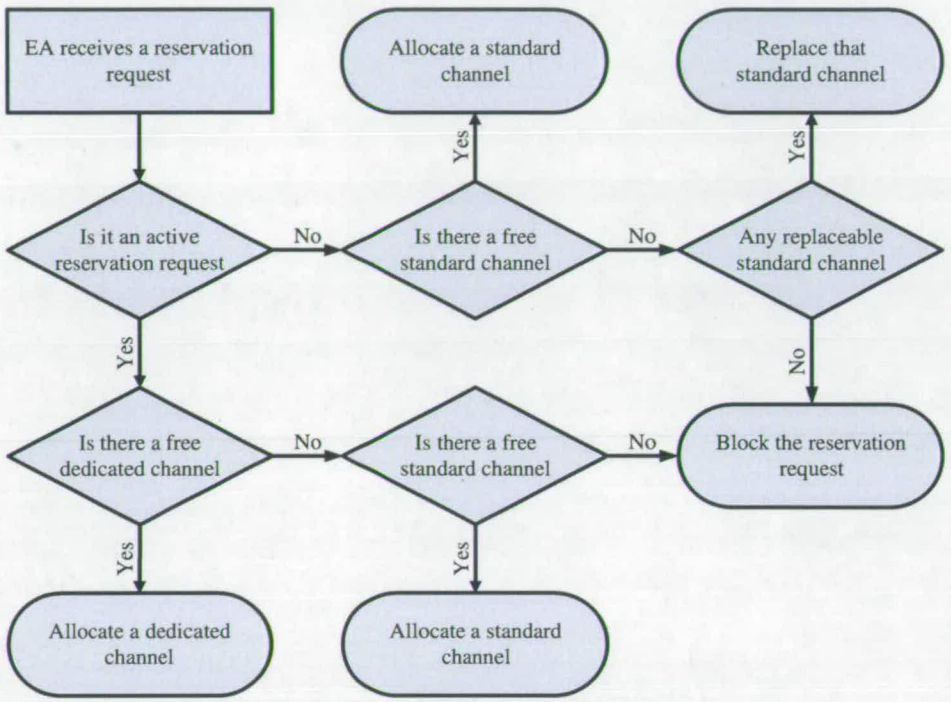


Fig. 1 The channel allocation procedure of the reservation optimised ARR scheme

parameter to determine which mobile node is more suitable for making a passive reservation path. This traffic class based admission control can be implemented in the *policy control* module [2] of RSVP.

Another important issue we should point out is that although the ARR schemes look similar to the handover prioritised schemes which are used in the cellular networks, they are significantly different in the ways in which resources are reserved. In the handover prioritised schemes, resources of a subnet are reserved for the mobile nodes in the neighbouring subnets and can be used by anyone that hands over into the subnet. On the other hand, in the ARR schemes, resources are reserved for specific mobile nodes and therefore the network resource utilisation is degraded in order to provide better QoS and mobility support.

3.4 Modularity

To improve the efficiency of the reservation optimised ARR scheme, position prediction algorithms, which determine in which neighbouring subnets a mobile node makes advance reservation paths, can be applied. With a precise position prediction algorithm and a low-cost signalling procedure (such as RSVP aggregation [24]), the signalling cost of the reservation optimised ARR scheme can be reduced.

Moreover, since the proposed scheme in fact consists of two admission control mechanisms, it can be easily integrated into existing ARR schemes by requiring them to implement these admission control mechanisms. As there are already agents in these ARR schemes, the only additional information required is the SMRs of the mobile nodes. Incidentally, collecting the session and mobility information of the mobile nodes is a basic requirement of future

next-generation communication networks [25]. In this way, the modularity of the proposed scheme is maintained.

3.5 Performance Evaluation

To compare the performance of different ARR schemes, we build continuous time Markov chain (CTMC) based analytical models. The analysis of resource management algorithms using CTMC at the state space level has been extensively studied in [26–28]. These approaches capture the state changes of the resources. However, specific patterns of transitions between states are required to produce a closed-form solution [28]. Since in the ARR schemes there are transitions between the active and passive states, such requirement is not met and so deriving a closed-form solution is very difficult. Therefore, in this work we use the performance modelling formalism PEPA. PEPA is chosen because firstly its component structure directly reflects the system structure, thereby providing a clear description of the system it models. Secondly, since PEPA is a process algebra language, it is quicker and easier to construct models than working directly at the state space level. Thirdly, since PEPA models can be solved numerically, some restrictions, which other modelling approaches such as queueing networks must follow to exhibit a product form solution, do not constrain PEPA models. Last but not least, sophisticated tools [29] have been developed which make both steady state and transient analysis of PEPA models convenient. A detailed discussion on the performance of the different ARR schemes is given in Sect. 6.

4 Performance Evaluation Process Algebra

4.1 Syntax of PEPA

Process algebras have been designed as formal description techniques for concurrent systems—systems that consist of subsystems interacting with each other. In process algebras, a *process* or an *agent* can perform *actions*, and a system is modelled from the perspective of its *behaviour* as interactions between processes. By using some basic axioms similar to those in elementary algebra[3], equational reasoning can be carried out in order to decide whether two systems are behaviourally equivalent, or to verify that a system satisfies a certain property, or to investigate other aspects of a system. Classical process algebras such as Calculus of Communicating Systems (CCS) [30] and Communicating Sequential Processes (CSP) [31] are designed for qualitative rather than quantitative analysis of a system. Performance Evaluation Process Algebra (PEPA) [32] is a timed and stochastic extension of classical process algebras that can be used for performance modelling of computer and communication systems. PEPA is a compositional approach that decomposes the system into subsystems that are smaller and more easily modelled. In PEPA a system is usually modelled to be composed of a group of *components* that engage in *activities*. This abstract description of a system is similar to the design process of a system and facilitates model construction. Generally, components model the physical or logical elements of a system and activities characterise the behaviour of these components. For example, a printing system can be considered to consist of a *Queue* component which buffers jobs and a *Printer* component which prints jobs.

[3] The simplest examples are $a + b = b + a$, $a + (b + c) = (a + b) + c$, etc. Here, the operator "+" denotes the *choice* operation rather than the *addition* operation, which will be explained later.

Each activity a in PEPA is defined as a pair (α, r)—the action type α, which can be regarded as the name of the activity, and the activity rate r, which is the parameter of an exponentially distributed random variable and specifies the duration of the activity. If a component P behaves as P' after completing activity a, then we can regard this behaviour as a component changing from state P to state P', through transition (α, r).

The PEPA formalism provides a small set of operators which are able to express the individual activities of components as well as the interactions between them. We only present the operators we used in our model in this subsection. For more details about PEPA operators, see [32].

Prefix: $(\alpha, r).P$

This component has a designated first activity which is of action type (or name) α and has a duration that is exponentially distributed with rate r, which gives a mean time of $1/r$. After completing this activity, the component $(\alpha, r).P$ behaves as P. For example, the *Printer* component in the above example can print a job and then suspend for a while before it is ready to take the next job. This behaviour can be expressed as:

$$\begin{aligned} Printer &\stackrel{def}{=} (print, r_1).Printer' \\ Printer' &\stackrel{def}{=} (suspend, r_2).Printer \end{aligned}$$

Choice: $P + Q$

This component may either behave as P or Q. All the enabled activities in P and Q are enabled in this component and compete with each other. The first activity to be completed will be an activity of P or Q and this will distinguish which component wins the race. When the first activity is completed, all the other activities will be abandoned. For example, let the component $Queue_i$ denotes the behaviour of the *Queue* component when there are i jobs in the queue. It can either allow another job to arrive (when the queue is not full) or have one of its jobs printed (when the queue is not empty). The *Queue* component can then be defined as $(i = 1, 2, \ldots, N-1)$:

$$\begin{aligned} Queue_0 &\stackrel{def}{=} (arrive, r_3).Queue_1 \\ Queue_i &\stackrel{def}{=} (arrive, r_3).Queue_{i+1} \\ &\quad + (print, \top).Queue_{i-1} \\ Queue_N &\stackrel{def}{=} (print, \top).Queue_{N-1} \end{aligned}$$

where N is the maximum size of the queue. The symbol "$\top$" means the rate of the activity is outside control of the component. In this example, the *Queue* component is *passive* with respect to the activity *print* since it cannot influence the rate at which jobs are printed.

Cooperation: $P \underset{L}{\bowtie} Q$

This component represents the interaction between P and Q. The set L is called the *cooperation set* and denotes a set of action types that must be carried out by P and Q together. For all activities whose action type is included in set L, P and Q must cooperate to complete it. However, other activities of P and Q which have types that are not included in set L will proceed independently. The rate of the *shared* activity is determined by the rate of the slower participant and is the smaller of the two rates. When the cooperation set L is empty, the two components proceed concurrently without any interaction between them. A shorthand notation $P \| Q$ is used to represent $P \underset{\varnothing}{\bowtie} Q$, and the symbol "$\|$" is referred to as the *parallel*

operator. N parallel components P can be expressed as a group: $P[N]$. For example, the printing system may have two parallel queues that share only one printer, and each *Queue* component cooperates with the *Printer* component on the activity *print* individually. This can be expressed as:

$$(Queue_0 \| Queue_0) \underset{\{print\}}{\bowtie} Printer$$

Constant: $P \stackrel{def}{=} Q$

The constant operator "$\stackrel{def}{=}$" can be used to associate names with behaviour. Its usage to define single components have been shown in the above examples. Moreover, it can be used for *system definition* which specifies how the system is constructed from the defined components, i.e., how the components cooperate with each other so that they express the behaviour of the system. For example, the printing system with two queues and one printer can be given a name *System*, which is associated with the cooperation between the components *Queue* and *Printer*. That is:

$$System \stackrel{def}{=} (Queue_0 \| Queue_0) \underset{\{print\}}{\bowtie} Printer$$

where $Queue_0$ and *Printer* define the initial behaviour of the corresponding components.

The system states of a PEPA model are the *feasible* combinations of the state of each component of that model. The above printing system example with two queues and one printer has 32 system states if N is 3. For example, the state

$$(Queue_2 \| Queue_1) \underset{\{print\}}{\bowtie} Printer'$$

is one of them.

4.2 Deriving Performance Measures

For any PEPA model, an underlying stochastic process can be generated. We can associate a state with a component, and the transitions between states are defined by the activities between them. Since the duration of a transition in PEPA is exponentially distributed, it has been shown that the stochastic process underlying a PEPA model is a continuous time Markov chain (CTMC). By deriving the steady state probability vector of the CTMC, and with the help of the Markov reward models (MRMs) [33], performance measures such as utilisation and throughput can be derived. By measuring these measures, model verification and system optimisation are facilitated and automated.

5 PEPA Models of the Conventional, Passive Reservation Limited and Reservation Optimised ARR Schemes

In this section, we present the PEPA models of the conventional and the reservation optimised ARR schemes. Moreover, as a stepping stone between the two models, the PEPA model of the ARR scheme which only implements the passive reservation limited mechanism is also built. These models aim to represent how the resources of a subnet are consumed by the mobile nodes. The subnet under observation is called a local subnet, and a mobile node is called a local and a foreign mobile node when it is in and out of the local subnet, respectively.

5.1 Traffic and Mobility Models

Before the PEPA models are built, it is necessary to make assumptions about the traffic and mobility models of a mobile node. The traffic model of a mobile node usually consists of two parts: the session arrival rate and the session duration time. Although recent study suggests that the Internet traffic at the packet level exhibits the long-range dependence property, the Poisson process is still a good model of the session arrival behaviour of the mobile node [34,35]. As for the session duration, the traditional exponential distribution is used [26,36,37].

The mobility model describes the distribution of the residence time of a mobile node in a area. By using different assumptions about the speed, direction and area shape, various types of distributions can be derived. However, the handover behaviour of a mobile node also depends on the type of handover procedure [38] and thus the distributions based on contrived mobility pattern and area shape are not really practical [39]. Therefore, without any proved probability distribution, the exponential distribution is chosen to model the mobile node's residence time in a area.

5.2 PEPA Model of the Conventional ARR Scheme

In the conventional ARR scheme, a local mobile node requires an active reservation path in the local subnet, while a foreign mobile node requires a passive reservation path in the local subnet. The local subnet does not discriminate the active and passive reservation requests, and the type of a reservation path is changed according to the movement of its reserver. There are two types of PEPA components in this model. The component *MN* models the behaviour of a mobile node, and a channel of the local subnet is modelled by the component *CHAN*.

5.2.1 Mobile Node

Since the resources of the local subnet are used by the local and foreign mobile nodes in different ways, the states of a mobile node are distinguished according to its position. Superscripts L and F are used to denote that the mobile node is in the local and neighbouring subnets, respectively. The mobile node is initially in the idle state MN^L_{Idle} (or state MN^F_{Idle}) and can move between different subnets (activities $move_{outwards}$ and $move_{inwards}$). Its mean sojourn times in local and neighbouring subnets are $1/\upsilon_{out}$ and $1/\upsilon_{in}$, respectively. The mobile node generates new session requests (activity *session_arrive*) at the rate of λ. Depending on the position of the mobile node, i.e., in states $MN^L_{Request}$ and $MN^F_{Request}$, it requires an active or a passive reservation path in the local subnet, respectively (activities $reserve_{active}$ and $reserve_{passive}$). If the request is blocked, the mobile node has to wait for the resources to become available, or during this waiting time it may move out of its current subnet and continue to require resources. If the request is admitted, the mobile node can start its session which has a mean duration of $1/\mu$.[4] At the engaged states $MN^L_{Engaged}$ and $MN^F_{Engaged}$, the mobile node actively and passively uses its reservation path, respectively (activities $session_{active}$ and $session_{passive}$). Since the foreign mobile node has already reserved resources in advance, it can continue its session when it hands over into the local subnet (activity $handover_{inwards}$). As for the local mobile node, it still occupies its reservation path in the local subnet when it hands over out

[4] In fact, a foreign mobile node cannot start its communication unless its active reservation request is admitted by its current subnet. To simplify the models, it is assumed that the foreign mobile node's active reservation requests are always admitted by its current subnet.

of the local subnet (activity $handover_{outwards}$). When the mobile node finishes its session, it returns to the idle state. The component MN is defined as:

$$\begin{aligned}
MN^L_{Idle} &\stackrel{def}{=} (session_arrive, \lambda).MN^L_{Request} \\
&\quad + (move_{outwards}, \upsilon_{out}).MN^F_{Idle} \\
MN^L_{Request} &\stackrel{def}{=} (reserve_{active}, r_{active}).MN^L_{Engaged} \\
&\quad + (move_{outwards}, \upsilon_{out}).MN^F_{Request} \\
MN^L_{Engaged} &\stackrel{def}{=} (session_{active}, \mu).MN^L_{Idle} \\
&\quad + (handover_{outwards}, \upsilon_{out}).MN^F_{Engaged} \\
MN^F_{Idle} &\stackrel{def}{=} (session_arrive, \lambda).MN^F_{Request} \\
&\quad + (move_{inwards}, \upsilon_{in}).MN^L_{Idle} \\
MN^F_{Request} &\stackrel{def}{=} (reserve_{passive}, r_{passive}).MN^F_{Engaged} \\
&\quad + (move_{inwards}, \upsilon_{in}).MN^L_{Request} \\
MN^F_{Engaged} &\stackrel{def}{=} (session_{passive}, \mu).MN^F_{Idle} \\
&\quad + (handover_{inwards}, \upsilon_{in}).MN^L_{Engaged}
\end{aligned}$$

5.2.2 Channel

A channel of the local subnet has three states: idle, active and passive. At state $CHAN_{Idle}$ the channel can accept active and passive reservation requests and go to the states $CHAN_{Active}$ and $CHAN_{Passive}$, respectively. The type of a reservation path is changed according to the movement of its reserver. When a mobile node finishes it session, its reservation path is released. The component $CHAN$ is defined as:

$$\begin{aligned}
CHAN_{Idle} &\stackrel{def}{=} (reserve_{active}, \top).CHAN_{Active} \\
&\quad + (reserve_{passive}, \top).CHAN_{Passive} \\
CHAN_{Active} &\stackrel{def}{=} (session_{active}, \top).CHAN_{Idle} \\
&\quad + (handover_{outwards}, \top).CHAN_{Passive} \\
CHAN_{Passive} &\stackrel{def}{=} (session_{passive}, \top).CHAN_{Idle} \\
&\quad + (handover_{inwards}, \top).CHAN_{Active}
\end{aligned}$$

5.2.3 System Definition

The component $CHAN$ synchronises with the component MN on the reservation request activities $reserve_{active}$ and $reserve_{passive}$, and the resource holding activities $reserve_{active}$ and $reserve_{passive}$. Moreover, the type of a reservation path is changed according to the mobile node's handover activities $handover_{outwards}$ and $handover_{inwards}$. The PEPA model of the conventional ARR scheme which consists of K mobile nodes and X channels is constructed as:

$$ARR^{CON} \stackrel{def}{=} \left(MN^L_{Idle}[K]\right) \underset{L}{\bowtie} \left(CHAN_{Idle}[X]\right)$$

where

$$L = \{reserve_{active}, reserve_{passive}, session_{active}, session_{passive}, handover_{outwards}, handover_{inwards}\}.$$

5.3 PEPA Model of the Passive Reservation Limited ARR Scheme

In the passive reservation limited ARR scheme, dedicated channels are set aside especially for active reservation paths. Therefore, there are two PEPA components $CHAN^S$ and $CHAN^D$ to model the standard and dedicated channels, respectively. The behaviour of the mobile node also needs to be modified to implement the passive reservation limited mechanism described in Sect. 3.1.

5.3.1 Mobile Node

Since a local mobile node can use both standard and dedicated channels, at state $MN^L_{Request}$ the local mobile node can either request a dedicated channel (activity $reserve_{prior}$), or request a standard channel when there is no free dedicated channel (activity $reserve_{active}$). Moreover, since the passive reservation limited scheme does not allow an active reservation path to change to the passive state, the engaged local mobile node has to request a passive reservation path when it hands over out of the local subnet, i.e., it goes to state $MN^F_{Request}$ instead of state $MN^F_{Engaged}$. The definitions of the other states and the behaviour of the component MN are the same as in the conventional ARR scheme model. The component MN is defined as:

$$
\begin{aligned}
MN^L_{Idle} &\stackrel{def}{=} (session\_arrive, \lambda).MN^L_{Request} \\
&\quad + (move_{outwards}, \upsilon_{out}).MN^F_{Idle} \\
MN^L_{Request} &\stackrel{def}{=} (reserve_{active}, r_{active}).MN^L_{Engaged} \\
&\quad + (reserve_{prior}, r_{prior}).MN^L_{Engaged} \\
&\quad + (move_{outwards}, \upsilon_{out}).MN^F_{Request} \\
MN^L_{Engaged} &\stackrel{def}{=} (session_{active}, \mu).MN^L_{Idle} \\
&\quad + (handover_{outwards}, \upsilon_{out}).MN^F_{Request} \\
MN^F_{Idle} &\stackrel{def}{=} (session\_arrive, \lambda).MN^F_{Request} \\
&\quad + (move_{inwards}, \upsilon_{in}).MN^L_{Idle} \\
MN^F_{Request} &\stackrel{def}{=} (reserve_{passive}, r_{passive}).MN^F_{Engaged} \\
&\quad + (move_{inwards}, \upsilon_{in}).MN^L_{Request} \\
MN^F_{Engaged} &\stackrel{def}{=} (session_{passive}, \mu).MN^F_{Idle} \\
&\quad + (handover_{inwards}, \upsilon_{in}).MN^L_{Engaged}
\end{aligned}
$$

5.3.2 Standard Channel

An idle standard channel can accept a passive reservation request from a foreign mobile node (activity $reserve_{passive}$) and is passively reserved. When the foreign mobile node hands over into the local subnet, the passively reserved standard channel becomes active, i.e., goes to state $CHAN^S_{Active}$. On the other hand, an idle standard channel can also be actively reserved by a local mobile node when there is no free dedicated channel (activity $reserve_{active}$). However, an actively reserved standard channel is released, i.e., goes to state $CHAN^S_{Idle}$, when the local mobile node hands over out of the local subnet. An engaged standard channel is released when its reserver finishes its session. The component $CHAN^S$ is defined as:

$$
\begin{aligned}
CHAN^S_{Idle} &\stackrel{def}{=} (reserve_{active}, \top).CHAN^S_{Active} \\
&\quad + (reserve_{passive}, \top).CHAN^S_{Passive} \\
CHAN^S_{Active} &\stackrel{def}{=} (session_{active}, \top).CHAN^S_{Idle} \\
&\quad + (handover_{outwards}, \top).CHAN^S_{Idle} \\
CHAN^S_{Passive} &\stackrel{def}{=} (session_{passive}, \top).CHAN^S_{Idle} \\
&\quad + (handover_{inwards}, \top).CHAN^S_{Active}
\end{aligned}
$$

5.3.3 Dedicated Channel

An idle dedicated channel can only accept active reservation requests from a local mobile node (activity $reserve_{prior}$) and be actively reserved. An engaged dedicated channel is released when its reserver finishes its session or hands over out of the local subnet. To guarantee that the local mobile node chooses the dedicated channels before the standard channels, the activity $reserve_{active}$ is defined as a self-transition at state $CHAN^D_{Active}$ and all the components $CHAN^D$ are required to cooperate on it. In this way, reserving a standard channel is only enabled when all the dedicated channels are engaged. The component $CHAN^D$ is defined as:

$$
\begin{aligned}
CHAN^D_{Idle} &\stackrel{def}{=} (reserve_{prior}, \top).CHAN^D_{Active} \\
CHAN^D_{Active} &\stackrel{def}{=} (session_{active}, \top).CHAN^D_{Idle} \\
&\quad + (handover_{outwards}, \top).CHAN^D_{Idle} \\
&\quad + (active\_reserve, \top).CHAN^D_{Active}
\end{aligned}
$$

5.3.4 System Definition

As in the conventional ARR scheme model, the component MN synchronises with the components $CHAN^S$ and $CHAN^D$ on the reservation request, resource holding and handover activities. Moreover, all the components $CHAN^D$ synchronise with each other and with the components $CHAN^S$ on the activity $reserve_{active}$, which guarantees that the dedicated channels are selected first. The PEPA model of the passive reservation limited ARR scheme which consists of K mobile nodes, X standard channels and Y dedicated channels is constructed as:

$$
ARR^{PRL} \stackrel{def}{=} \left(MN^L_{Idle}[K]\right) \underset{L_1}{\bowtie} \left(\left(CHAN^S_{Idle}[X]\right) \underset{L_2}{\bowtie} \left(\underbrace{CHAN^D_{Idle} \underset{L_2}{\bowtie} CHAN^D_{Idle} \cdots \underset{L_2}{\bowtie} CHAN^D_{Idle}}_{Y} \right) \right)
$$

where

$$
L_1 = \{reserve_{active}, reserve_{prior}, reserve_{passive}, session_{active}, session_{passive}, handover_{outwards}, handover_{inwards}\}, \quad \text{and} \quad L_2 = \{reserve_{active}\}.
$$

5.4 PEPA Model of the Reservation Optimised ARR Scheme

The reservation optimised ARR scheme includes the passive reservation limited and the SMR-based replacement mechanisms. The objective of the SMR-based replacement mechanism is to make the best usage of the standard channels in a subnet. Only the foreign mobile nodes with the highest SMR are eligible to make passive reservation paths in the local subnet. Since the SMR of a mobile node is dynamic and the replacement procedure is hard to implement using the performance modelling formalisms, an approach that achieves equivalent effects as the replacement procedure is employed. In this approach, the mobile nodes in the network are classified into two groups: the fast mobile nodes and the slow mobile nodes. This type of classification does not lose generality since there will always be some mobile nodes that have higher SMR than the others and are eligible to request passive reservation paths. These mobile nodes can be regarded as the fast mobile nodes and the rest can be regarded as the slow mobile nodes. The fast and slow mobile nodes are modelled by the components *Fast_MN* and *Slow_MN* respectively.

5.4.1 Fast Mobile Node

A fast mobile node can make both active and passive reservation requests. Its states and behaviour are the same as the component *MN* in the passive reservation limited ARR scheme model. The component *Fast_MN* is defined as:

$$
\begin{aligned}
Fast\_MN^{L}_{Idle} &\stackrel{def}{=} (session\_arrive, \lambda).Fast\_MN^{L}_{Request} \\
&\quad + (move_{outwards}, \upsilon_{out}).Fast\_MN^{F}_{Idle} \\
Fast\_MN^{L}_{Request} &\stackrel{def}{=} (reserve_{active}, r_{active}).Fast\_MN^{L}_{Engaged} \\
&\quad + (reserve_{prior}, r_{prior}).Fast\_MN^{L}_{Engaged} \\
&\quad + (move_{outwards}, \upsilon_{out}).Fast\_MN^{F}_{Request} \\
Fast\_MN^{L}_{Engaged} &\stackrel{def}{=} (session_{active}, \mu).Fast\_MN^{L}_{Idle} \\
&\quad + (handover_{outwards}, \upsilon_{out}).Fast\_MN^{F}_{Request} \\
Fast\_MN^{F}_{Idle} &\stackrel{def}{=} (session\_arrive, \lambda).Fast\_MN^{F}_{Request} \\
&\quad + (move_{inwards}, \upsilon_{in}).Fast\_MN^{L}_{Idle} \\
Fast\_MN^{F}_{Request} &\stackrel{def}{=} (reserve_{passive}, r_{passive}).Fast\_MN^{F}_{Engaged} \\
&\quad + (move_{inwards}, \upsilon_{in}).Fast\_MN^{L}_{Request} \\
Fast\_MN^{F}_{Engaged} &\stackrel{def}{=} (session_{passive}, \mu).Fast\_MN^{F}_{Idle} \\
&\quad + (handover_{inwards}, \upsilon_{in}).Fast\_MN^{L}_{Engaged}
\end{aligned}
$$

5.4.2 Slow Mobile Node

A slow mobile node behaves differently from a fast mobile node when it is in the neighbouring subnets. The passive reservation requests of the slow foreign mobile node are always blocked by the local subnet (state $Slow\_MN^{F}_{Blocked}$), and it only stops requesting until it finishes its session[5] or moves into the local subnet and requests an active reservation path. Moreover, since the slow mobile node in the neighbouring subnets has no effect on the resource utilisation of the local subnet, at state $Slow\_MN^{F}_{Blocked}$ its session and handover behaviour are named differently (activities *session* and *handover*) so that they do not synchronise with the

[5] Remember it is assumed that a foreign mobile node is always admitted by its current subnet.

channels of the local subnet. Note that if the slow mobile node hands over into the local subnet before the end of its session, it needs to request an active reservation path since it does not have resources reserved in advance. The component *Slow_MN* is defined as:

$$
\begin{aligned}
Slow\_MN^{L}_{Idle} &\overset{def}{=} (session\_arrive, \lambda).Slow\_MN^{L}_{Request} \\
&\quad + (move_{outwards}, \upsilon^{slow}_{out}).Slow\_MN^{F}_{Idle} \\
Slow\_MN^{L}_{Request} &\overset{def}{=} (reserve_{active}, r_{active}).Slow\_MN^{L}_{Engaged} \\
&\quad + (reserve_{prior}, r_{prior}).Slow\_MN^{L}_{Engaged} \\
&\quad + (move_{outwards}, \upsilon^{slow}_{out}).Slow\_MN^{F}_{Blocked} \\
Slow\_MN^{L}_{Engaged} &\overset{def}{=} (session_{active}, \mu).Slow\_MN^{L}_{Idle} \\
&\quad + (handover_{outwards}, \upsilon^{slow}_{out}).Slow\_MN^{F}_{Blocked} \\
Slow\_MN^{F}_{Idle} &\overset{def}{=} (session\_arrive, \lambda).Slow\_MN^{F}_{Blocked} \\
&\quad + (move_{inwards}, \upsilon^{slow}_{in}).Slow\_MN^{L}_{Idle} \\
Slow\_MN^{F}_{Blocked} &\overset{def}{=} (session, \mu).Slow\_MN^{F}_{Idle} \\
&\quad + (handover, \upsilon^{slow}_{in}).Slow\_MN^{L}_{Request}
\end{aligned}
$$

5.4.3 Standard Channel and Dedicated Channel

The definitions of the components $CHAN^S$ and $CHAN^D$ are the same as those in the passive reservation limited ARR scheme model and they are omitted here.

5.4.4 System Definition

The cooperation relations between the mobile nodes and the channels in this model are the same as those in the passive reservation limited ARR scheme model. Moreover, the number of standard channels is the maximum number of mobile nodes that are eligible to make passive reservation paths, and thus is also the number of the mobile nodes that can be regarded as the fast mobile nodes. Therefore, the PEPA model of the reservation optimised ARR scheme which consists of X fast mobile nodes, Z slow mobile nodes, X standard channels and Y dedicated channels is constructed as:

$$
ARR^{RO} \overset{def}{=} \left(\left(Fast\_MN^{L}_{Idle}[X] \right) \parallel \left(Slow\_MN^{L}_{Idle}[Z] \right) \right) \underset{L_1}{\bowtie} \left(\left(CHAN^{S}_{Idle}[X] \right) \underset{L_2}{\bowtie} \left(\underbrace{CHAN^{D}_{Idle} \underset{L_2}{\bowtie} CHAN^{D}_{Idle} \cdots \underset{L_2}{\bowtie} CHAN^{D}_{Idle}}_{Y} \right) \right)
$$

where

$$
L_1 = \{reserve_{active}, reserve_{prior}, reserve_{passive}, session_{active}, session_{passive}, handover_{outwards}, handover_{inwards}\}, \quad \text{and} \quad L_2 = \{reserve_{active}\}.
$$

Table 1 Parameters settings of the PEPA models of the conventional, passive reservation limited, and reservation optimised ARR schemes

Rate	Corresponding activities	Value (s^{-1})
λ	$session\_arrive$	$[30:30:450]^{-1}$
μ	$session_{active}$, $session_{passive}$, $session$	$[45:45:675]^{-1}$
υ_{out}	$move_{outwards}$, $handover_{outwards}$	480^{-1}
υ_{out}^{slow}	$move_{outwards}$, $handover_{outwards}$	960^{-1}
υ_{in}	$move_{inwards}$, $handover_{inwards}$	960^{-1}
υ_{in}^{slow}	$move_{inwards}$, $handover_{inwards}$, $handover$	1920^{-1}
r_{active}	$reserve_{active}$	0.1^{-1}
r_{prior}	$reserve_{prior}$	0.1^{-1}
$r_{passive}$	$reserve_{passive}$	0.2^{-1}

6 Performance Evaluation

Since the fundamental goal of the work presented in this paper is to investigate how the resources could be managed when they are under-provisioned, the effects of traffic intensity on the blocking probabilities of active and passive reservation requests are evaluated. These performance measures are of interest because they reflect the network congestion levels for different types of reservation paths. Moreover, the mean numbers of active and passive reservation paths in a subnet are also evaluated to investigate the effect of different ARR schemes on the resource utilisation of a subnet.

To guarantee the models are numerically tractable, all the models have four mobile nodes and three channels. In the passive reservation limited ARR scheme model there are two standard channels and one dedicated channel. In the reservation optimised ARR scheme model there are two fast mobile nodes, two slow mobile nodes, two standard channels and one dedicated channel. Since the mobile nodes in all the PEPA model are individually expressed, performance measures can be derived by observing either a single mobile node or all the mobile nodes as a whole, and the former is chosen in the evaluation. The first *MN* component in the conventional and passive reservation limited models, and the first *Fast_MN* and *Slow_MN* in the reservation optimised model are chosen to be investigated.

6.1 Parameter Settings

The traffic intensity can be tuned by the session arrival interval $1/\lambda$ and the session duration $1/\mu$, and their variation ranges are listed in Table 1. The mean residence times of a (fast) mobile node within and outside the local subnet are set to 480 and 960 s, respectively. As for a slow mobile node, its sojourn time in an area is twice as long as that of its fast counterpart. The mean delay of the active reservation request messages is set to 0.1 s, and since the passive reservation requests are sent from the neighbouring subnets, their mean delay is set to 0.2 s. The rates of all the activities are listed in Table 1.

6.2 Active Reservation Blocking Probability

The active reservation blocking happens in the system states in which the investigated mobile node is in the local subnet and requires an active reservation path (state $MN^L_{Request}$, $Fast\_MN^L_{Request}$ and $Slow\_MN^L_{Request}$ in respective models) while no channel is in idle state.

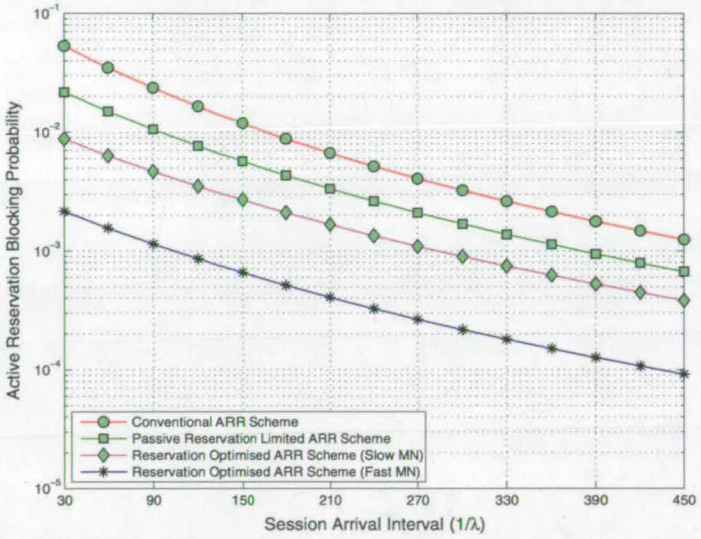


Fig. 2 The effect of session arrival interval on active reservation blocking probability

Therefore, this probability can be calculated by summing up the probabilities of the system states that follow the above specification.

Figure 2 shows the impact of session arrival interval on the active reservation blocking probability for the three schemes. The results are calculated with the mean session duration set to 360 s, and the y-axis is in the logarithmic scale. It can be observed from the figure that the active reservation blocking probability decreases when the session arrives less frequently. The passive reservation limited ARR scheme has a lower blocking probability (from 2.17×10^{-2} to 6.69×10^{-4}) than the conventional ARR scheme (from 5.32×10^{-2} to 1.24×10^{-3}), because it sets aside dedicated channels for active reservation requests. In the reservation optimised ARR scheme, since fewer foreign mobile nodes are eligible to request standard channels for making passive reservation paths, the competition for the resources in the local subnet is less severe. Therefore, both fast and slow mobile nodes in the reservation optimised ARR scheme have a lower active reservation blocking probability than the other two ARR schemes. Moreover, as a slow mobile node stays longer in the local subnet than a fast mobile node, at the same session arrival interval, it generates more requests during its sojourn time in the local subnet. Therefore, the slow mobile node is more likely to be rejected (from 8.78×10^{-3} to 3.82×10^{-4}) than the fast mobile node (from 2.16×10^{-3} to 9.18×10^{-5}) for active reservation requests.

A similar improvement on the active reservation blocking probability in the passive reservation limited and reservation optimised schemes can also be observed in Fig. 3, where the session arrival interval set to 240 s and the session duration is changed. It is clear that the engaged mobile nodes hold the resources for longer time when the session duration grows, and thus the active reservation blocking probability increases. When the session duration is less than 90 s, the passive reservation limited ARR scheme and the conventional ARR scheme have close performance on the active reservation blocking. However, as the session duration gets larger, the former grows from 1.27×10^{-5} to 5.84×10^{-3} and clearly outperforms the latter which increases from 1.57×10^{-5} to 1.50×10^{-2}. Again, the reservation optimised ARR scheme has the lowest active reservation blocking probability, in which the

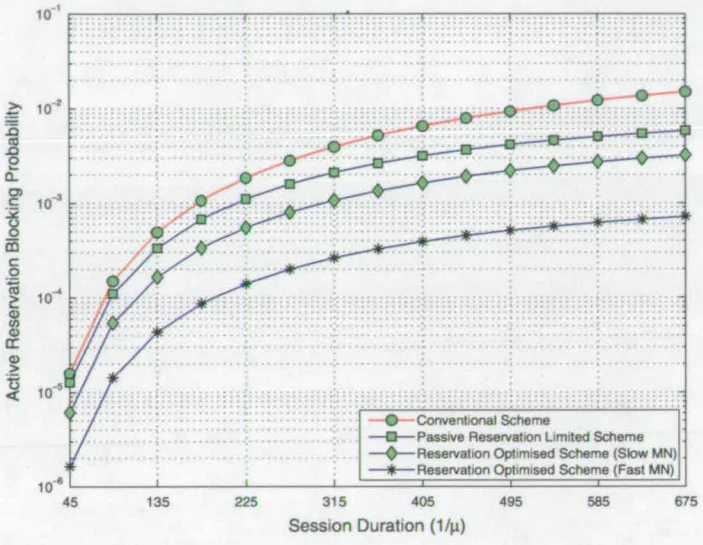


Fig. 3 The effect of session duration on active reservation blocking probability

slow mobile node has a blocking probability growing from 6.13×10^{-6} to 3.22×10^{-3} and the fast mobile node has a lower blocking probability ranging from 1.64×10^{-6} to 7.21×10^{-4}.

6.3 Passive Reservation Blocking Probability

Similarly to the active reservation blocking probability, the passive reservation blocking happens in the system states in which the mobile node is in the neighbouring subnets and requires a passive reservation path in the local subnet (state $MN^{F}_{Request}$ and $Fast\_MN^{F}_{Request}$ in respective models) but no channel is in idle state. As for the slow mobile node in the reservation optimised ARR scheme, since it is always blocked when it is out of the local subnet, its passive reservation blocking probability is not investigated.

Figure 4 shows how passive reservation requests of the mobile nodes are affected by the restrictions on the passive reservations. The conventional ARR scheme has a passive reservation blocking probability ranging from 1.05×10^{-1} to 2.48×10^{-3}. Since the passive reservation limited ARR scheme restricts the resource for passive reservations, it has a higher blocking probability than the conventional ARR scheme which decreases from 2.58×10^{-1} to 3.07×10^{-2}. An interesting observation is that the fast mobile node in the reservation optimised ARR scheme has a passive reservation blocking probability that ranges from 3.69×10^{-2} to 3.44×10^{-3} and is smaller than that in the conventional ARR scheme when the session arrival interval is small. Remember that in the reservation optimised ARR scheme, the resource competition is less severe since not all the foreign mobile nodes are allowed to make passive reservation paths. The results indicate that at high traffic intensities, the resource competition caused by the number of the mobile nodes has greater effect than that caused by limited resource. In other words, reducing the number of requests can compensate the resource restriction. However, this difference decreases and at session arrival intervals larger than 270 s, the fast mobile node experiences larger blocking probability in

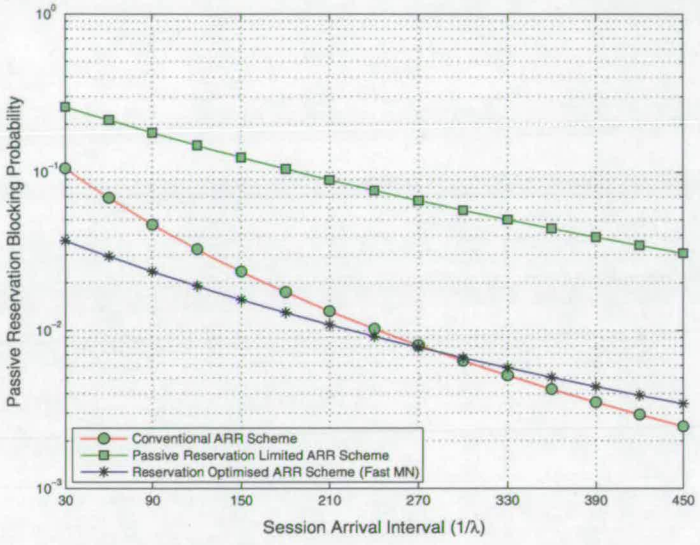


Fig. 4 The effect of session arrival interval on passive reservation blocking probability

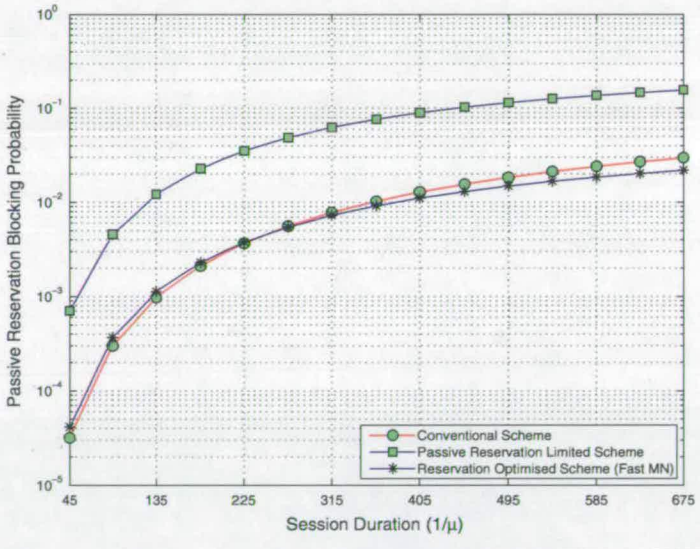


Fig. 5 The effect of session duration on passive reservation blocking probability

the reservation optimised ARR scheme, which indicates that the limited resource now has a larger effect on the passive reservation blocking probability.

Similar results can be observed in Fig. 5 where the session duration is the tuning parameter. The passive reservation limited ARR scheme has the highest passive reservation blocking probability that increases from 7.06×10^{-4} to 1.57×10^{-1}. The fast mobile node in the reservation optimised ARR scheme has a passive reservation blocking probability ranging

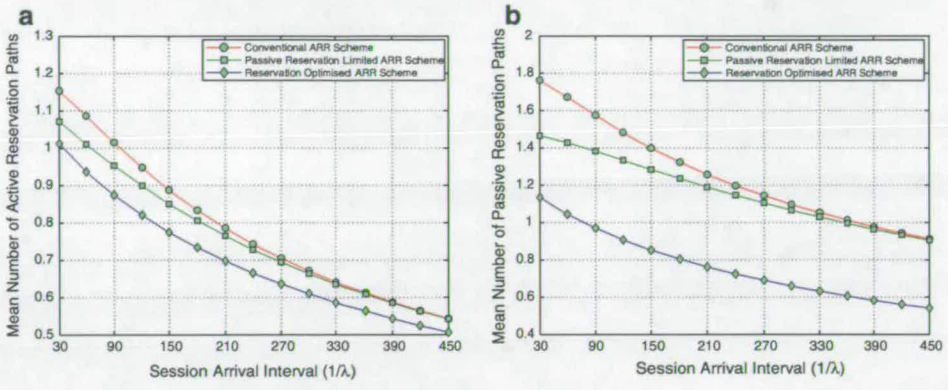


Fig. 6 The effect of session arrival interval on mean numbers of active and passive reservation paths. **a** Mean number of active reservation paths. **b** Mean number passive reservation paths

from 4.15×10^{-5} to 2.20×10^{-2}, which is very close to that in the conventional ARR scheme ranging from 3.18×10^{-5} to 2.98×10^{-2}. However, the former outperforms the latter at higher traffic intensities, i.e. when the session duration is larger than 315 s.

6.4 Mean Numbers of Active and Passive Reservation Paths

To study how different ARR schemes affect the resource utilisation of a subnet, the mean numbers of the active and passive reservation paths in a subnet are also investigated. The mean number of a certain type of reservation paths in a model can be computed as the weighted sum of the numbers of that type of reservation paths in all the system states, where the weights are the equilibrium probabilities of the system states.

Figure 6a shows the effect of session arrival interval on the mean number of active reservation paths in a subnet. An interesting observation is that the conventional ARR scheme has the largest number of active reservation paths in a subnet, while the reservation optimised ARR scheme ranks the last. This indicates that although the passive reservation limited ARR scheme and the reservation optimised ARR scheme give a higher priority to the active reservation requests, this does not result in a higher proportion of active reservation paths in the subnet. In fact, the mean number of active reservation paths could also be affected by how long the resources are actively engaged by the mobile nodes. At the same traffic intensity, since in the reservation optimised ARR scheme the competition between the active reservation requests for resources is less severe than the other ARR schemes, its active resource engagement time is shorter.

Similar trends can be seen in Fig. 6b, the passive reservation limited ARR scheme has a smaller number of passive reservation paths than the conventional ARR scheme because of the restriction on the available resources for passive reservation paths. The reservation optimised ARR scheme reduces this number further by preventing slow mobile nodes from making passive reservation paths.

Figure 7a and b show the relationship between the mean numbers of active and passive reservation paths in a subnet and the session duration. Again, the conventional ARR scheme has the largest numbers of both active and passive reservation paths, followed by the passive reservation limited and then the reservation optimised ARR schemes. However, the results of the conventional and the passive reservation limited ARR schemes are almost the same at session durations less than 225 s.

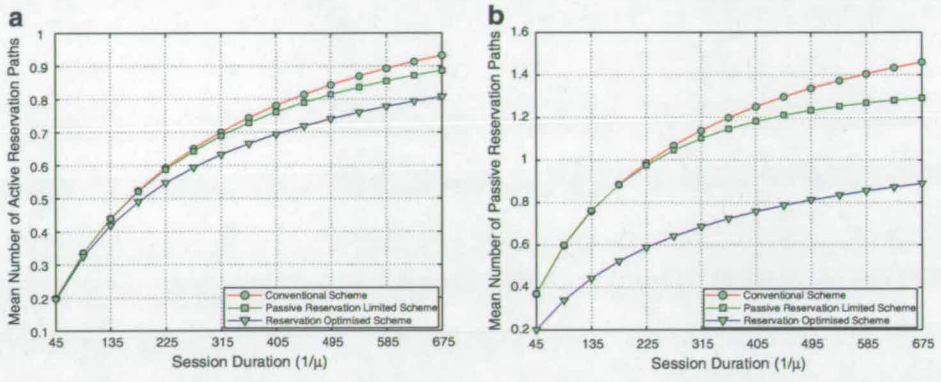


Fig. 7 The effect of session duration on mean numbers of active and passive reservation paths. **a** Mean number of active reservation paths. **b** Mean number passive reservation paths

6.5 Discussion

By restricting some of the foreign mobile nodes from making passive reservations, the reservation optimised ARR scheme achieves a better utilisation of the network resources and reduces both active and passive reservation blocking probability effectively. Although the performance improvement is achieved at the expense of introducing handover interruption to the slow mobile nodes, the proposed scheme is still reasonable since:

1. Passive reservation blocking only means that a foreign mobile node cannot make an advance reservation path in the local subnet. Although there could be an interruption when this foreign mobile node hands over into the local subnet, this type of reservation blocking has no effect on its current QoS session. On the other hand, an active reservation request implies that there is a local mobile node which really needs the requested resources to start its communication. Therefore, it is practical to give a higher priority to the active reservation requests.
2. When a foreign mobile node that is not granted a passive reservation path hands over into the local subnet, it becomes a local mobile node and requires an active reservation path. Since the proposed scheme sets aside dedicated channels for active reservation paths and reduces the active reservation blocking probability dramatically, this new local mobile node is highly likely to acquire an active reservation path and continue its communication.

7 Conclusion

In this paper, we propose a novel reservation optimised ARR scheme. This scheme aims to balance the number of active and passive reservation paths in a subnet. Our motivation is that the passive reservation paths in a subnet are not actively used by their owners and thus they waste the network resources from the perspective of the QoS traffic. The results indicate that the proposed reservation optimised ARR scheme achieves a better utilisation of the network resources and balances the active and passive reservation blocking probability effectively. This is achieved by setting aside dedicated channels for active reservation paths and restricting some of the foreign mobile nodes from making passive reservation paths.

Although the performance improvement is achieved at the expense of introducing handover interruption to the slow mobile nodes, the proposed scheme is still reasonable.

Acknowledgements The work reported in this paper has formed part of the Ubiquitous Services Core Research Programme of the Virtual Centre of Excellence in Mobile & Personal Communications, Mobile VCE, www.mobilevce.com. This research has been funded by the Industrial Companies who are Members of Mobile VCE, with additional financial support from the UK Governments Technology Strategy Board (previously DTI). Fully detailed technical reports on this research are available to Industrial Members of Mobile VCE. J. Hillston is also supported by EPSRC Advanced Research Fellowship EP/c543696/01 and EU FET-IST Global Computing 2 project SENSORIA (Software Engineering for Service-Oriented Overlay Computers (IST-3-016004-IP-09)). H. Wang and D. Laurenson acknowledge the support of the Scottish Funding Council for the Joint Research Institute with the Heriot-Watt University which is a part of the Edinburgh Research Partnership.

References

1. Braden, R., Clark, D., & Shenker, S. (1994, June). Integrated services in the internet architecture: An overview, RFC 1633.
2. Braden, R., Zhang, L., Berson, S., Herzog, S., & Jamin, S. (1997, September). Resource ReSerVation Protocol (RSVP)—Version 1 functional specification, RFC 2205.
3. Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., & Weiss, W. (1998, December). An architecture for differentiated service, RFC 2475.
4. Lo, S.-C., Lee, G., Chen, W.-T., & Liu, J.-C. (2004). Architecture for mobility and QoS support in all-IP wireless networks. *IEEE Journal on Selected Areas in Communications, 22*(4), 691–705.
5. Mankin, A., Baker, F., Braden, B., Bradner, S., O'Dell, M., & Romanow, A. et al. (1997, September). Resource ReSerVation Protocol (RSVP)—Version 1 applicability statement some guidelines on deployment, RFC 2208.
6. Bernet, Y. (2000). The complementary roles of RSVP and differentiated services in the full-service QoS network. *IEEE Communications Magazine, 38*(2), 154–162.
7. Moon, B., & Aghvami, H. (2001). RSVP extensions for real-time services in wireless mobile networks. *IEEE Communications Magazine, 39*(12), 52–59.
8. Johnson, D., Perkins, C., & Arkko, J. (2004, June). Mobility support in IPv6, RFC 3775.
9. Soliman, H., Castelluccia, C., Malki, K. E., & Bellier, L. (2005, August). Hierarchical Mobile IPv6 Mobility Management (HMIPv6), RFC 4140.
10. Manner, J., Toledo, A. L., Mihailovic, A., Munoz, H. L. V., Hepworth, E., & Khouaja, Y. (2002). Evaluation of mobility and quality of service interaction. *Computer Networks, 38*(2), 137–163.
11. Leu, S.-J., & Chang, R.-S. (2003). Integrated service mobile internet: RSVP over mobile IPv4&6. *Mobile Networks and Applications, 8*(6), 635–642.
12. Chaskar, H., (2003, September). Requirements of a quality of service (QoS) solution for mobile IP, RFC 3583.
13. Kuo, G.-S., & Ko, P.-C. (2003). Dynamic RSVP protocol. *Wireless Networks, 41*(5), 130–135.
14. Tseng, C.-C., Lee, G.-C., Liu, R.-S., & Wang, T.-P. (2003). HMRSVP: a hierarchical mobile RSVP protocol. *Wireless Networks, 9*(2), 95–102.
15. Paskalis, S., Kaloxylos, A., Zervas, E., & Merakos, L. (2003). An efficient RSVP-mobile IP interworking scheme. *Mobile Networks and Applications, 8*(3), 197–207.
16. Chen, W.-T., & Huang, L.-C. (2000). RSVP mobility support: A signaling protocol for integrated services internet with mobile hosts. In *Proceedings IEEE international conference on computer communications '00* (pp. 1283–1292).
17. Talukdar, A. K., Badrinath, B. R., & Acharya, A. (2001). MRSVP: A resource reservation protocol for an integrated services network with mobile hosts. *Wireless Networks, 7*(1), 5–19.
18. Awduche, D., & Agu, E. (1997). Mobile extensions to RSVP. In *Proceedings IEEE international conference on computer communications and networks '97* (pp. 132–136).
19. Levine, D., Akyildiz, I., & Naghshineh, M. (1997). A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept. *IEEE/ACM Transactions on Networking, 5*(1), 1–12.
20. Huang, N.-F., & Chen, W.-E. (2003). RSVP extensions for real-time services in hierarchical mobile IPv6. *Mobile Networks and Applications, 8*(6), 625–634.
21. Huston, G. (2000, November). Next steps for the IP QoS architecture, RFC 2990.

22. Pack, S., Shen, X. S., Mark, J. W., & Pan, J. (2007). Adaptive route optimization in hierarchical mobile IPv6 networks. *IEEE Transactions on Mobile Computing, 6*(8), 903–914
23. Pack, S., Kwon, T., & Choi, Y. (2004). A mobility-based load control scheme at mobility anchor point in hierarchical mobile IPv6 networks. In *Proceedings IEEE Global Telecommunications Conference '04* (pp. 3431–3435).
24. Baker, F., Iturralde, C., Faucheur, F., & Davie, B. (2001, September). Aggregation of RSVP for IPv4 and IPv6 reservations. RFC 3175.
25. Kibria, M., & Jamalipour, A. (2007). On designing issues of the next generation mobile network. *IEEE Network, 21*(1), 6–13.
26. Hong, D., & Rappaport, S. (1986). Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures. *IEEE Transactions on Vehicular Technology, 35*(3), 77–92.
27. Yeung, K., & Nanda, S. (1996). Channel management in Microcell/Macrocell cellular radio systems. *IEEE Transactions on Vehicular Technology, 45*(4), 601–612.
28. Fang, Y., & Zhang, Y. (2002). Call admission control schemes and performance analysis in wireless mobile networks. *IEEE Transactions on Vehicular Technology, 51*(2), 371–382.
29. PEPA Tools. Online. Available http://www.dcs.ed.ac.uk/pepa/tools/.
30. Milner, R. (1989). *Communication and concurrency*. Upper Saddle River: Prentice Hall.
31. Hoare, C. A. R. (1985). *Communicating sequential processes*. Upper Saddle River: Prentice Hall.
32. Hillston, J. (1996). *A compositional approach to performance modelling*. Cambridge: Cambridge University Press.
33. Bolch, G., Greiner, S., de Meer, H., & Trivedi, K. S. (1998). *Queueing networks and Markov chains: Modeling and performance evaluation with computer science applications*. New York: Wiley.
34. Lam, D., Cox, D., & Widom, J. (1997). Teletraffic modeling for personal communications services. *IEEE Communications Magazine, 35*(2), 79–87.
35. Karagiannis, T., Molle, M., & Faloutsos, M. (2004). Long-range dependence ten years of internet traffic modeling. *IEEE Internet Computing, 8*(5), 57–64.
36. Fang, Y., Chlamtac, I., & Lin, Y.-B. (1997). Modeling PCS networks under general call holding time and cell residence time distributions. *IEEE/ACM Transactions on Networking, 5*(6), 893–906.
37. Lin, Y.-B. (1997). Modeling techniques for large-scale PCS networks. *IEEE Communications Magazine, 35*(2), 102–107.
38. Nasser, N., Hasswa, A., & Hassanein, H. (2006). Handoffs in fourth generation heterogeneous networks. *IEEE Communications Magazine, 44*(10), 96–103.
39. Kourtis, S., & Tafazolli, R. (2000). Evaluation of handover related statistics and the applicability of mobility modelling in their prediction. In *Proceedings IEEE international symposium on personal, indoor and mobile radio communications '00* (pp. 665–670).

Author Biographies

Hao Wang received a B.Eng. degree in Information Engineering from the Southeast University, China, in 2003 and a M.Sc. degree in Signal Processing and Communications from The University of Edinburgh in 2005. He is currently working towards a Ph.D. degree in the School of Engineering at The University of Edinburgh. His research interests include developing and performance modelling of mobility and QoS management schemes in mobile and wireless networks.

David I. Laurenson obtained a B.Eng. degree in Computer Science and Electronics followed by a Ph.D. in Electronics and Electrical Engineering from The University of Edinburgh in 1990 and 1994, respectively. He was appointed as a Lecturer in 1994 with research interests in mobile radio communications, ranging from propagation issues to ad-hoc networking.

Jane Hillston is Professor of Quantitative Modelling in the School of Informatics at The University of Edinburgh and holds an Advanced Research Fellowship from the Engineering and Physical Sciences Research Council. Her principal research interests are in the use of stochastic process algebras to model and analyse computer systems and the development of efficient solution techniques for such models. Prof. Hillston received the B.A. and M.Sc. degrees in Mathematics from the University of York (UK) and Lehigh University (USA), respectively. After a brief period working in industry, she joined the Department of Computer Science at The University of Edinburgh, as a research assistant in 1989. She received the Ph.D. degree in Computer Science from that university in 1994. Her work on the stochastic process algebra PEPA was recognised by the British Computer Society in 2004 who awarded her the first Roger Needham Award.

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings.

Evaluation of RSVP and Mobility-aware RSVP Using Performance Evaluation Process Algebra

Hao Wang and David I. Laurenson
Institute for Digital Communications,
Joint Research Institute for Signal & Image Processing,
School of Engineering & Electronics,
The University of Edinburgh
Email: {H.Wang, Dave.Laurenson}@ed.ac.uk

Jane Hillston
Laboratory for Foundations of Computer Science,
School of Informatics,
The University of Edinburgh
Email: Jane.Hillston@ed.ac.uk

***Abstract*—As a resource reservation mechanism, the Resource ReSerVation Protocol (RSVP) faces a lot of challenges when applying it to the wireless and mobile networks. The interworking problems of RSVP and mobility management protocols have been extensively discussed over the last decade. As the solutions of this problem, mobility-aware RSVP schemes that integrate RSVP and micro-mobility management are becoming more and more popular. Therefore, the investigation on how much they improve the performance of the basic RSVP is necessary and useful. Instead of the traditional simulation based approaches, in this paper we introduce a formal performance evaluation formalism, named Performance Evaluation Process Algebra (PEPA), and employ it to investigate the performance of the basic RSVP and mobility-aware RSVP. Important performance metrics such as handover blocking probability and signalling cost are presented.**

I. Introduction

As many real-time services and multimedia applications become popular, providing guaranteed quality of service (QoS) to Internet users is an important issue for the next generation of traffic management. One of the proposed solutions is the Integrated Service [1] that utilises a signalling protocol such as Resource ReSerVation Protocol (RSVP) [2] to control end-to-end packet delay. However, due to the mobility of mobile users, RSVP becomes inefficient because there is a disruption of QoS traffic when a mobile node changes its point of attachment to the network. A lot of variants of the basic RSVP have been proposed and most of them tackle the problem from the perspective of the macro-mobility management protocols such as Mobile IP. Detailed surveys of RSVP over Mobile IP can be found in [3]–[5]. On the other hand, it is proposed in [6] that for every mobile node's movement to a new IP subnet, the micro-mobility management protocol is preferable to its global counterpart and a global-mobility management protocol is not even strictly required to provide node mobility. Moreover, as we will see in section II, a micro-mobility management protocol such as Hierarchical Mobile IPv6 (HMIPv6) [7] has inherent characteristics which facilitate the deployment of RSVP in a mobile environment. Therefore, schemes that integrate RSVP and micro-mobility management mechanisms have become widely accepted as the best approach to combining mobility and QoS, and it is necessary and useful to investigate their expected performance.

Most of the previous efforts on evaluating the enhancements of the mobility-aware RSVP are carried out by simulation. Specific network topologies and traffic scenarios are used in the simulations and performance metrics such as packet delay and throughput are obtained. However, simulation is not always a reliable means of determining performance metrics since the results are usually subject to the specific simulation setup. The contribution of our work is that we are the first to build Markovian models of both basic RSVP and mobility-aware RSVP to assess their performance. Furthermore, these models are built using a formal performance evaluation formalism, named Performance Evaluation Process Algebra (PEPA). From these PEPA models, we derive important performance metrics such as handover blocking probability and signalling cost, and demonstrate the advantages of the mobility-aware RSVP. Moreover, we should point out that our models are independent of the specific implementations of RSVP and mobility-aware RSVP schemes and capture the essential characteristics underlying them.

The rest of the paper is structured as follows. In Section II we introduce the basic RSVP and the mobility-aware RSVP schemes that integrate RSVP and micro-mobility management mechanisms. In Section III we give a short introduction to the PEPA formalism. In Section IV the PEPA models of both basic and mobility-aware RSVP are presented. The performance of the two RSVP schemes are analysed in Section V and we give our conclusion in Section VI.

II. RSVP and Mobility-aware RSVP

RSVP is a receiver-oriented resource reservation setup protocol for simplex data flows. It can be used by a host to request specific qualities of service from the network and by routers to establish and maintain the required QoS. Since in RSVP the resource reservation in a network is identified by the IP addresses of the communicating ends, one of the major incompatibilities between RSVP and mobility management when providing QoS guarantees in a mobile network is that the receivers must re-establish reservation whenever a mobile node performs a handover. This disruption during handover significantly degrades QoS-sensitive services. To reduce the resource re-establishment time, one of the solutions is to

978-1-4244-2075-9/08/$25.00 ©2008 IEEE

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings.

localise the reservation signalling within the affected part of the path in the network [8].

Previous work on deploying RSVP in a micro-mobility management enabled network [9]–[11] takes advantage of the two-layer care-of addresses of a mobile node. Here we take HMIPv6 as the example. In HMIPv6, there is a new mobility agent called Mobility Anchor Point (MAP) that covers a group of access routers (ARs). Every time a mobile node moves into a MAP domain, it acquires an on-link care-of address (LCoA) referring to the AR which it is connected to and a regional care-of address (RCoA) referring to the MAP. Outside the MAP domain, the mobile node is identified by its RCoA and all the packets addressed at RCoA are intercepted by the MAP. The MAP will then forward these packets to the mobile node at LCoA. Therefore, when the mobile node performs a handover within a MAP domain, i.e., switches to a new AR connecting to the same MAP, only the LCoA is changed and the RCoA remains the same. It then follows that a mobile node actually only needs to change the reservation path between the AR and the MAP, and maintains the same reservation path outside the MAP domain as long as it uses the same RCoA. In the proposed mobility-aware RSVP schemes, there is an agent (Mobility Proxy in [9] and QoS Agent in [10]) in the access network that assists the mobile node to make this kind of partial resource reservation. This agent can be located at the gateway of the access network. Every time the mobile node performs a handover, it notifies the agent of the current binding between its LCoA and RCoA. Upon receiving this information, the agent is capable of intercepting and looking into the RSVP messages and swapping the LCoA and RCoA in a way that the reservation below and above the agent is identified by the LCoA and the RCoA respectively. Therefore, as long as the mobile node moves within the same MAP domain, the RSVP signalling only traverses to the agent and the reservation re-establishment time is reduced. For details about the operation of these schemes, see [9], [10].

III. Performance Evaluation Process Algebra

The term *process algebras* refers to mathematical theories which model and reason about the structure and behaviour of a system in an algebraic framework. Performance Evaluation Process Algebra (PEPA) [12] is a timed and stochastic extension of classical process algebra such as CCS [13] and CPS [14] that can be used for performance modelling of computer and communication systems. PEPA is a compositional approach that decomposes the system into subsystems that are smaller and more easily modelled. In PEPA a system is usually composed of a group of *components* that engage in *activities*. Generally, components model the physical or logical elements of a system and activities characterise the behaviour of these components. Each activity a in PEPA is defined as a pair (α, r) — the action type α, which can be regarded as the name of the activity, and the activity rate r, which is an exponentially distributed random variable and specifies the duration of the activity. If a component P behaves as P' after completing activity a, then we can regard this behaviour as a component changing from state P to state P', through transition (α, r).

The PEPA formalism provides a small set of operators which are able to express the individual activities of components as well as the interactions between them. We only present the operators we used in our model in this section. For more details about PEPA operators, see [12].

Prefix: $(\alpha, r).P$

This component has a designated first activity which is of action type (or name) α and has a duration that is exponentially distributed with rate r, which gives a mean time of $1/r$. A larger rate implies a faster completion of an activity. After completing this activity, the component $(\alpha, r)\,.P$ behaves as P.

Choice: $P + Q$

This component may either behave as P or Q. All the enabled activities in P and Q are enabled in this component and compete with each other. The first activity to be completed will be an activity of P or Q and this will distinguish which component wins the race. When the first activity is completed, all the other activities will be abandoned. For example, the component $(\alpha, r_1).P' + (\beta, r_2).Q'$ is more likely to subsequently behave as P' if r_1 is larger than r_2.

Cooperation: $P \underset{L}{\bowtie} Q$

This component represents the interaction between P and Q. The set L is called the *cooperation set* and denotes a set of action types that must be carried out by P and Q together. For all activities whose action type is included in set L, P and Q must cooperate to complete it. However, other activities of P and Q which have types that are not included in set L will proceed independently. The rate of the *shared* activity is determined by the rate of the slower participant and is the smaller of the two rates. In PEPA an activity can have an unspecified rate making it a *passive activity* and its rate is labelled as $\top$. This means that although the component which has this passive activity is required to engage in the cooperation, it has no influence on the rate at all.

Parallel: $P \| Q$

This component represents two concurrent but completely independent components. This is simply a shorthand notation for $P \underset{\emptyset}{\bowtie} Q$.

Constant: $P \stackrel{def}{=} Q$

This expression is used to assign names to components.

System Definition:

Since PEPA is a compositional approach, in PEPA a system is described as an interaction of components. The system definition specifies how the system is constructed from the defined components.

To generate a stochastic process which represents the PEPA model, we can associate a state with a component, and the transitions between states are defined by the activities between them. Since the duration of the transition in PEPA is exponentially distributed, it has been shown that the stochastic process underlying a PEPA model is a discrete state space, continuous time Markov chain (CTMC). By deriving the steady state probability vector of the CTMC, and with the

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings.

help of the Markov reward models (MRMs) [15], performance measures such as utilisation and throughput can be derived. These measures can facilitate model verification and system optimisation.

IV. PEPA Models of Basic RSVP and Mobility-aware RSVP

In this work, we build CTMC-based analytical models using PEPA. PEPA is chosen because firstly its component structure directly reflects the system structure, thereby providing a clear description of the system it models. Secondly, since PEPA is a process algebra language, it is quicker and easier to construct models than working directly at the state space level. Thirdly, since PEPA models can be solved numerically, some restrictions, which other modelling approaches such as queueing networks must follow to exhibit a product form solution, do not constrain PEPA models. Last but not least, sophisticated tools [16] have been developed which make both steady state and transient analysis of PEPA models convenient. In this section, the PEPA models of the basic RSVP and the mobility-aware RSVP are presented. We remind the reader that our models are not restricted to any specific implementation of these schemes but are general models which capture the essential characteristics of basic RSVP and mobility-aware RSVP.

A. PEPA Model of Basic RSVP

The scenario used in our model is a mobile node moving within a local domain and communicating with other nodes. We refer to the networks below and above the merge point of the old and new RSVP path as the *lower network* and the *upper network* respectively. The lower network consists of the whole or part of the mobile node's access network and the upper network consists of the Internet core network which is usually heavily loaded. Here we borrow the concept of *channel* from cellular networks to represent the network resources. Therefore, there are three elementary types of PEPA components in the model, which are *Mobile Node* (*MN*), *Lower Network Channel* (*LNC*) and *Upper Network Channel* (*UNC*). The last two components represent the resources in the lower network and upper network respectively.

Mobile Node: The *MN* models the behaviour of a mobile node. The *MN* is initially in the idle state MN_0. It requests a reservation of both *LNC* and *UNC* (state MN_1) after receiving a call request which arrives at the rate of λ. If both *LNC* and *UNC* are available, the request is accepted and the *MN* can start its RSVP session (state MN_2). Otherwise, the request is blocked and the *MN* keeps requesting a reservation until it is finally allocated one (state MN_1). The average length of an RSVP session is assumed to be $1/\mu$. During this session, the *MN* can perform a localised handover at the rate of α, and then it needs to request a new reservation of both *LNC* and *UNC* in order to continue its session (state MN_3). (We assume the *MN* implements the *local repair* [2] option, so it can request a new reservation almost immediately after a handover.) After the session is finished, the *MN* tears down its current reservation (state MN_4). The component *MN* is defined as:

$$\begin{aligned}
MN_0 &\stackrel{def}{=} (call_arrive, \lambda).MN_1 \\
MN_1 &\stackrel{def}{=} (reserve_all, r).MN_2 \\
&\quad + (block, b).MN_1 \\
MN_2 &\stackrel{def}{=} (session, \mu).MN_4 \\
&\quad + (handover, \alpha).MN_3 \\
MN_3 &\stackrel{def}{=} (reserve_all, r).MN_2 \\
&\quad + (block, b).MN_3 \\
MN_4 &\stackrel{def}{=} (tear_all, t).MN_0
\end{aligned}$$

Lower Network Channel: The *LNC* component models the resources in the lower network. It can be reserved and torn down explicitly by a mobile node in a way similar to a queue. If the mobile node performs a handover, the old reservation of the mobile node expires after an average period of $1/\gamma$. (Note that the basic RSVP [2] only suggests a node explicitly tears down its old reservation at the end of an RSVP session.) When the *LNC* is fully engaged, it blocks the requests of the mobile nodes. If the capacity of the *LNC* is M, it is defined as:

$$\begin{aligned}
LNC_0 &\stackrel{def}{=} (reserve_all, \top).LNC_1 \\
LNC_i &\stackrel{def}{=} (reserve_all, \top).LNC_{i+1} \\
&\quad + (tear_all, \top).LNC_{i-1} \\
&\quad + (expire, \gamma).LNC_{i-1} \quad (\forall\, i \in [1, M-1]) \\
LNC_M &\stackrel{def}{=} (block, \top).LNC_M \\
&\quad + (tear_all, \top).LNC_{M-1} \\
&\quad + (expire, \gamma).LNC_{M-1}
\end{aligned}$$

Upper Network Channel: The *UNC* component models the resources in the upper network and its behaviour is the same as *LNC*. If the capacity of the *UNC* is N, it is defined as:

$$\begin{aligned}
UNC_0 &\stackrel{def}{=} (reserve_all, \top).UNC_1 \\
UNC_i &\stackrel{def}{=} (reserve_all, \top).UNC_{i+1} \\
&\quad + (tear_all, \top).UNC_{i-1} \\
&\quad + (expire, \gamma).UNC_{i-1} \quad (\forall\, i \in [1, N-1]) \\
UNC_N &\stackrel{def}{=} (block, \top).UNC_N \\
&\quad + (tear_all, \top).UNC_{N-1} \\
&\quad + (expire, \gamma).UNC_{N-1}
\end{aligned}$$

Channel Monitor: The *CM* component is an assistant component in our model. Its function is to guarantee that the *expire* activity is only performed after a *handover* (by requiring a cooperation on it between *CM* , *LNC* and *UNC*) and the number of performed *expire* and *handover* activities are the same. It is defined as:

$$\begin{aligned}
CM_0 &\stackrel{def}{=} (handover, \top).CM_1 \\
CM_i &\stackrel{def}{=} (handover, \top).CM_{i+1} \\
&\quad + (expire, \top).CM_{i-1} \quad (\forall\, i \in [1, M-1]) \\
CM_M &\stackrel{def}{=} (expire, \top).CM_{M-1}
\end{aligned}$$

System Definition: Since in basic RSVP the mobile node reserves and releases resources in both lower and upper network at the same time, the activity $reserve\_all$ and $tear\_all$

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings.

must be carried out by *MN*, *LNC* and *UNC* together. Either *LNC* or *UNC* can cooperate with the *MN* on the *block* activity when they are fully engaged. The *CM* synchronises with *MN* on the *handover* activity and with *LNC* and *UNC* on the *expire* activity. In this way, the *expire* activity can only be carried out after a *handover*. To guarantee the built model is numerically tractable while keeping the generality, in our model there are 3 parallel mobile nodes and M and N are set to be 5 and 3 respectively. Therefore, the RSVP model is constructed as:

$$System \stackrel{def}{=} (MN_0 \| MN_0 \| MN_0) \underset{L_1}{\bowtie} \Big((UNC_0 \underset{L_2}{\bowtie} LNC_0) \underset{L_3}{\bowtie} CM_0 \Big)$$

where

$$L_1 = \{reserve\_all, tear\_all, block, handover\},$$

$$L_2 = \{reserve\_all, tear\_all, expire\},\ L_3 = \{expire\}.$$

B. PEPA Model of Mobility-aware RSVP

In the mobility-aware RSVP, the *MN* only requests a new reservation in the lower network after a handover. Since a PEPA component essentially describes the behaviour of an entity, we can simply modify the *MN* component so that when it is in state MN_3 it performs $reserve\_lnc$ instead of $reserve\_all$. The activity $reserve\_lnc$ represents a reservation request for lower network resource only. The component *MN* is modified as:

$$\begin{array}{lcl} MN_0 & \stackrel{def}{=} & (call\_arrive, \lambda).MN_1 \\ MN_1 & \stackrel{def}{=} & (reserve\_all, r).MN_2 \\ & + & (block, b).MN_1 \\ MN_2 & \stackrel{def}{=} & (session, \mu).MN_4 \\ & + & (handover, \alpha).MN_3 \\ MN_3 & \stackrel{def}{=} & (reserve\_lnc, r).MN_2 \\ & + & (block, b).MN_3 \\ MN_4 & \stackrel{def}{=} & (tear\_all, t).MN_0 \end{array}$$

The *LNC* component also needs to be modified so that it is aware of the new type of request which only asks for reservation in the lower network. To make it a fair comparison, in our model the *MN* does not explicitly remove the old reservation after a handover as required in the proposed mobility-aware RSVP schemes. The component *LNC* is modified as:

$$\begin{array}{lcl} LNC_0 & \stackrel{def}{=} & (reserve\_all, \top).LNC_1 \\ & + & (reserve\_lnc, \top).LNC_1 \\ LNC_i & \stackrel{def}{=} & (reserve\_all, \top).LNC_{i+1} \\ & + & (reserve\_lnc, \top).LNC_{i+1} \\ & + & (tear\_all, \top).LNC_{i-1} \\ & + & (expire, \gamma).LNC_{i-1}\ (\forall\, i \in [1, M-1]) \\ LNC_M & \stackrel{def}{=} & (block, \top).LNC_M \\ & + & (tear\_all, \top).LNC_{M-1} \\ & + & (expire, \gamma).LNC_{M-1} \end{array}$$

Accordingly, since there is no need for a new upper network reservation after a handover, an upper network reservation is established and torn down at the start and the end of an RSVP session. It never expires because it is always active during an RSVP session. The component *UNC* is modified as:

$$\begin{array}{lcl} UNC_0 & \stackrel{def}{=} & (reserve\_all, \top).UNC_1 \\ UNC_i & \stackrel{def}{=} & (reserve\_all, \top).UNC_{i+1} \\ & + & (tear\_all, \top).UNC_{i-1}\ (\forall\, i \in [1, N-1]) \\ UNC_N & \stackrel{def}{=} & (block, \top).UNC_N \\ & + & (tear\_all, \top).UNC_{N-1} \end{array}$$

System Definition: The system definition of mobility-aware RSVP is the same as the basic RSVP model except for the cooperation sets L_1 and L_2. The L_1 now includes the $reserve\_lnc$ activity, and the $expire$ activity is removed from L_2 since the upper network reservation does not expire.

$$System \stackrel{def}{=} (MN_0 \| MN_0 \| MN_0) \underset{L_1}{\bowtie} \Big((UNC_0 \underset{L_2}{\bowtie} LNC_0) \underset{L_3}{\bowtie} CM_0 \Big)$$

where

$$L_1 = \{reserve\_all, reserve\_lnc, tear\_all, block, handover\},$$

$$L_2 = \{reserve\_all, tear\_all\},\ L_3 = \{expire\}.$$

TABLE I
PARAMETERS VALUES

Type (Role)	Average Time (sec.)	Rate (1/sec.)
λ (call arrival interval)	10-100	0.01-0.1
μ (session duration)	180	1/180
α (handover interval)	120-600	1/600-1/120
γ (soft state lifetime)	90	1/90
r (reserve signalling)	0.1	10
b (block signalling)	0.1	10
t (tear signalling)	0.1	10

V. PERFORMANCE EVALUATION

Since the network between the two communicating ends usually consists of the Internet core network where the traffic is highly congested, an optimum utilisation of it is both practically and economically required. A more congested network usually results in a higher handover blocking probability and a larger signalling delay implies a longer interruption of QoS sensitive traffic. The mobility-aware RSVP schemes are especially designed to eliminate the unnecessary consumptions of the network resources and reduce signalling overhead. Therefore, the performance measures we investigate are the probability that the mobile nodes are rejected for continuing their session after handover and the signalling cost of both basic RSVP and mobility-aware RSVP. Before deriving these metrics, we first need to set the activity rates within the model. We make the traditional assumption that the call arrival interval, session duration and handover interval are exponentially distributed. We assume the average lifetime of an RSVP soft state is 90 seconds as suggested in [2]. For the RSVP signalling such as requesting and blocking, they are set to be 0.1 second. These activity rates are shown in Table I.

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings.

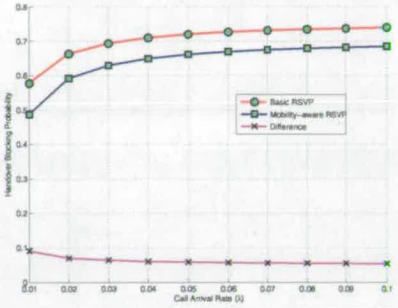


Fig. 1. Handover Blocking Probability vs. Call Arrival Rate

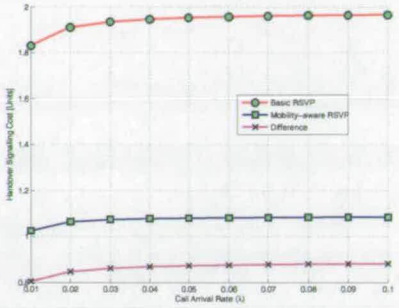


Fig. 3. Handover Signalling Cost vs. Call Arrival Rate

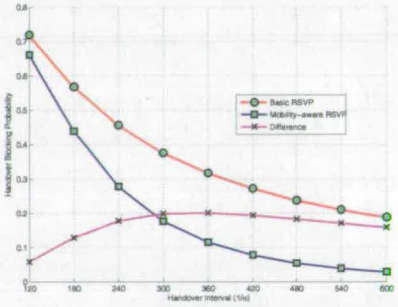


Fig. 2. Handover Blocking Probability vs. Handover Interval

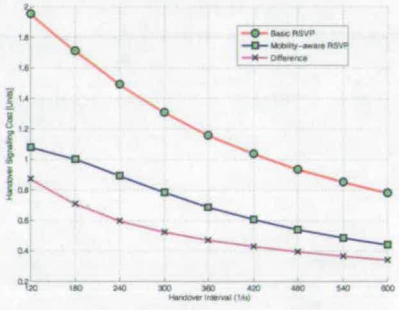


Fig. 4. Handover Signalling Cost vs. Handover Interval

A. Handover Blocking Probability

To derive the handover blocking probability, we just need to calculate the probabilities of the states in which the mobile nodes request reservations after handover (*MN* in state MN_3) but the lower or upper network is fully engaged (*LNC* in state LNC_5 or *UNC* in state UNC_3). Fig. 1 shows the effects of the call arrival rate on the handover blocking probability. The handover interval, i.e., the residence time of a mobile node in a subnet, is set to be 120 seconds. It can be observed that the blocking probability increases as expected for both basic RSVP and mobility-aware RSVP and their performance gets closer as the arrival rate of RSVP sessions grows. However, since the mobility-aware RSVP does not require a new reservation in the upper network after a handover, it has a lower blocking probability. We should point out that the reason why the blocking probability is so high is because in our model the network capacity is relatively much smaller than the number of mobile nodes, and we do it particularly to emphasise the congestion of the network and highlight the benefits of the mobility-aware RSVP.

The impact of the mobile node's mobility is also investigated, as shown in Fig. 2. The call arrival rate is set to be 0.05. It is easy to see that when the handover frequency decreases, the reservation requirements for network resources are reduced and thereby a lower handover blocking probability. Moreover, the mobility-aware RSVP has a much lower blocking probability compared to the basic RSVP when the handover interval is around 360 seconds and the difference between them gets smaller when the mobile nodes slow down. We can also observe that the performance of the two schemes gets close at small handover intervals, and this is because the network is overcongested and the mobility-aware RSVP does not improve the performance very much. Therefore, it can be concluded that in most typical scenarios, mobility-aware RSVP is less affected by the mobile node's mobility.

B. Handover Signalling Cost

Since one of the major benefits of the mobility-aware RSVP is reducing the scope which the RSVP signalling messages traverse after a handover, another performance measure of interest is the handover signalling cost. By employing the Markov reward model (MRM) [15] on a CTMC, we can easily compute the signalling costs associated with the two schemes. MRMs have been widely used in Markov decision theory to assign rewards (or costs) to states of Markov processes for system optimisation [17]. In our models, we associate rewards with the activities of interest, then the reward associated with a

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the ICC 2008 proceedings.

state is calculated by summing up the rewards of the activities that the state enables. If r_i is the reward associated with state S_i, and $\pi(\cdot)$ is the steady state probability vector of the CTMC, then the total reward R is

$$R = \sum_i r_i * \pi(S_i).$$

To derive the handover signalling cost, the activities we investigate are $reserve\_lnc$ and $reserve\_all$ after a handover, i.e. when *MN* is in state MN_3. We assign the costs of one unit and two units to $reserve\_lnc$ and $reserve\_all$ respectively. (That is, we assume the cost of sending a basic RSVP signalling message is two times that of a mobility-aware RSVP signalling message.) The effect of the call arrival rate on the signalling cost is depicted in Fig. 3. The handover interval is set to be 120 seconds. We can find that as the call arrival rate grows, the signalling costs for both RSVP schemes only increases a little at the beginning and then remain almost unchanged. This is mainly because we only take account of the signalling after a handover. Although the mobile nodes generate RSVP sessions more frequently at a larger call arrival rate, the handovers take place during an RSVP session and thus the associated cost is not sensitive to the call arrival rate. Therefore, the handover signalling cost is mostly dependent on the mobility of the mobile nodes, as shown in Fig. 4. The mobility-aware RSVP experiences a lower signalling cost than the basic RSVP since the former restricts the signalling within the affected area of the network. For the large handover intervals, the difference between the two schemes gets smaller because the mobile nodes seldom change their points of attachment and the benefits of the mobility-aware RSVP is less apparent. This again shows that the mobility-aware RSVP is more suitable in a mobile environment.

VI. Conclusion

Since RSVP and mobility management protocols were designed independently, the efficient integration of them is necessary to provide a QoS guaranteed mobility to the mobile node. Several mobility-aware RSVP schemes are proposed and it is necessary and practical to investigate how much they improve the basic RSVP. Instead of the traditional simulation based approaches, in this paper, we build Markovian models of both basic and mobility-aware RSVP schemes to evaluate their performance. Moreover, these models are built using a formal performance modelling formalism named PEPA. The PEPA models are built in a general way and so they are independent of the specific implementations of the schemes. Owing to PEPA's component structure, these models exhibit clear representations of the mechanisms underlying the proposed schemes. We investigate the impacts of the call arrival rate and handover interval on the probability of being blocked and the signalling cost after a handover. The results indicate that the mobility-aware RSVP outperforms the basic RSVP on both handover blocking probability and signalling cost as expected. These enhancements are achieved by avoiding unnecessary resource reservation in the unaffected part of the network and limiting RSVP signalling to the lower network. In our future work, we will investigate other problems of combining mobility and QoS, such as efficient resource pre-reservation schemes, in the PEPA framework.

Acknowledgement

The work reported in this paper has formed part of the Ubiquitous Services Core Research Programme of the Virtual Centre of Excellence in Mobile & Personal Communications, Mobile VCE, www.mobilevce.com. This research has been funded by the Industrial Companies who are Members of Mobile VCE, with additional financial support from the UK Governments Technology Strategy Board (previously DTI). Fully detailed technical reports on this research are available to Industrial Members of Mobile VCE.

J. Hillston is also supported by EPSRC Advanced Research Fellowship EP/c543696/01 and EU FET-IST Global Computing 2 project SENSORIA (Software Engineering for Service-Oriented Overlay Computers (IST-3-016004-IP-09)). H. Wang and D. Laurenson acknowledge the support of the Scottish Funding Council for the Joint Research Institute with the Heriot-Watt University which is a part of the Edinburgh Research Partnership.

References

[1] R. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: an Overview," RFC 1633, Jun. 1994.
[2] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification," RFC 2205, Sep. 1997.
[3] B. Moon and H. Aghvami, "RSVP extensions for real-time services in wireless mobile networks," *IEEE Commun. Mag.*, vol. 39, no. 12, pp. 52–59, 2001.
[4] J. Manner, A. L. Toledo, A. Mihailovic, H. L. V. Munoz, E. Hepworth, and Y. Khouaja, "Evaluation of mobility and quality of service interaction," *Comput. Netw.*, vol. 38, no. 2, pp. 137–163, 2002.
[5] S.-J. Leu and R.-S. Chang, "Integrated service mobile Internet: RSVP over mobile IPv4&6," *Mob. Netw. Appl.*, vol. 8, no. 6, pp. 635–642, 2003.
[6] J. Kempf, "Problem Statement for Network-Based Localized Mobility Management (NETLMM)," RFC 4830, Apr. 2007.
[7] H. Soliman, C. Castelluccia, K. E. Malki, and L. Bellier, "Hierarchical Mobile IPv6 Mobility Management (HMIPv6)," RFC 4140, Aug. 2005.
[8] H. Chaskar, "Requirements of a Quality of Service (QoS) Solution for Mobile IP," RFC 3583, Sep. 2003.
[9] S. Paskalis, A. Kaloxylos, E. Zervas, and L. Merakos, "An efficient RSVP-mobile IP interworking scheme," *Mob. Netw. Appl.*, vol. 8, no. 3, pp. 197–207, 2003.
[10] N.-F. Huang and W.-E. Chen, "RSVP extensions for real-time services in hierarchical mobile IPv6," *Mob. Netw. Appl.*, vol. 8, no. 6, pp. 625–634, 2003.
[11] H.-W. Ferng, W.-Y. Kao, J.-J. Huang, and D. Shiung, "A dynamic resource reservation scheme designed for improving multicast protocols in HMIPv6-based networks," in *Proc. Vehicular Technology Conference-Spring '06*, 2006, pp. 961–965.
[12] J. Hillston, *A Compositional Approach to Performance Modelling*. Cambridge University Press, 1996.
[13] R. Milner, *Communication and Concurrency*. Prentice Hall, 1989.
[14] C. A. R. Hoare, *Communicating Sequential Processes*. Prentice Hall, 1985.
[15] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. John Wiley & Sons, 1998.
[16] PEPA Tools. [Online]. Available: http://www.dcs.ed.ac.uk/pepa/tools/
[17] R. A. Howard, *Dynamic Probabilistic Systems, Volume 2: Semi-Markov and Decision Processes*. John Wiley & Sons, 1971.

An SMR Based Advance Resource Reservation Scheme For Combined Mobility and QoS Provisioning

Hao Wang and David I. Laurenson
Institute for Digital Communications,
Joint Research Institute for Signal & Image Processing,
School of Engineering & Electronics,
The University of Edinburgh
Email: {H.Wang, Dave.Laurenson}@ed.ac.uk

Jane Hillston
Laboratory for Foundations of Computer Science,
School of Informatics,
The University of Edinburgh
Email: Jane.Hillston@ed.ac.uk

***Abstract*—One of the major problems of deploying RSVP in the mobile environment is called the *advance resource reservation* problem. If an RSVP reservation path is reserved in advance in the subnet that a mobile node will visit, the mobile node can continue its QoS session smoothly when it hands over to that subnet. However, if too many network resources are used for advance reservation, new QoS sessions originating from that subnet will experience a higher probability of being blocked needlessly. In this paper, we propose a new advance resource reservation scheme that properly constrains the amount of advance reservations in a subnet and only allows the mobile nodes with large value of session-to-mobility ratio (SMR) to make advance reservation. We evaluate the proposed scheme and the results show that our scheme can effectively reduce both active and passive reservation blocking probabilities and achieves a better utilisation of the network resources, especially when the traffic intensity is high.**

I. Introduction

As many real-time services and multimedia applications become popular, providing guaranteed quality of service (QoS) to Internet users is an important issue for the next generation of traffic management. One of the proposed solutions is the Integrated Service [1] that utilises a signalling protocol such as Resource ReSerVation Protocol (RSVP) [2] to control end-to-end packet delay. RSVP is a receiver-oriented resource reservation setup protocol for simplex data flows. It can be used by a host to request specific amount of resources from the network and by routers to establish and maintain the required QoS. However, resource reservation using conventional RSVP exhibits a lot of deficiencies in the mobile environment. This is because in RSVP the resource reservation is identified by the IP addresses of the communicating ends, and the mobile node must re-establish the reservation after a layer-3 handover. Generally, when the mobile node changes the data flow path after handover, the congestion level along the path is also changed [3]. If the new path is overcongested, the available bandwidth along the new path may not be sufficient to satisfy the requirements of the handovered QoS session. To solve this problem, it has been suggested that the required resources be reserved in advance in the subnets that a mobile node may visit. Once the resources are guaranteed before handover, a mobile node can continue its QoS session smoothly after it switches its connectivity to another subnet. However, making advance reservations in a subnet is a waste of network resources from the QoS traffic's point of view and it also increases the blocking probability of the reservation requests originating from the mobile nodes in that subnet. In this paper, we propose a session-to-mobility ratio (SMR) based advance resource reservation scheme. This scheme combines advance resource reservation and call admission control (CAC) mechanisms and achieves a better network resource utilisation. The fundamental purpose of our scheme is to restrict the amount of advance reserved resources in a subnet and only allow the most eligible mobile nodes to make advance reservation.

The rest of the paper is structured as follows. In Section II we review the most representative advance resource reservation schemes. In Section III we present the procedure of our SMR based advance resource reservation scheme. The performance of our scheme is discussed in Section IV and we give our conclusions in Section V.

II. Related Work

Advance resource reservation approaches aim to make resource reservation in advance before a mobile node actually performs a handover. Previous proposals can be classified into two types: agent-based approaches and multicast-based approaches.

- **Agent-based approaches:** In the agent-based approaches, there are two types of reservations: *active* and *passive*. A mobile node makes an active reservation from its current subnet and passive reservations from other subnets to the correspondent node. When the mobile node moves out of its current subnet, the passive reservation in the newly visited subnet is switched to the active state and the active reservation along the old path is changed to the passive state. In every subnet there is a Proxy Agent (PA) [4] (or Mobility Agent (MA) [5]) which is in charge

978-1-4244-2644-7/08/$25.00 ©2008 IEEE

of the resource reservation process. A mobile node needs to inform the PAs of its neighbouring subnets which it may visit of its reservation information and require them to make passive reservations for it.

- **Multicast-based approaches:** In the multicast-based approaches, RSVP messages and ordinary data packets are delivered to a mobile node using IP multicast routing. As in the agent-based approaches, there is a Mobile Proxy (MP) [6] (or QoS Agent (QA) [7]) in every subnet. Before the mobile node starts a reservation, the mobile node and the MPs in its neighbouring subnets join a multicast group. In this way, a handover of the mobile node can be modelled as leaving and joining the branches of a multicast tree. The local MP makes a *Conventional Reservation* between the two communicating ends and the neighbour MPs make *Predictive Reservations* on behalf of the mobile node. When the mobile node moves to a new subnet, the reservation states of the old and new paths are changed accordingly. Therefore, the mobile node can continue its QoS session without interruption when it moves out of its current subnet.

With the help of advance resource reservation approaches, the handover dropping probability of a QoS session is reduced. However, unlike the concept of a *guard channel* in the cellular networks, an advance reservation in a subnet is made exclusively by a mobile node and is not actively used by its reserver. Moreover, allowing too many advance reservations in a subnet will increase the blocking probability of the reservation requests originating from the mobile nodes in that subnet, which reduces the Grade of Service (GoS) [8] of that subnet. The approaches that allow traffic with lower QoS requirements to temporarily use the passively reserved but unused resources [4]–[7] is not reliable, since the resources borrowed by a QoS session have to be returned when their reservers reclaim them. On the other hand, only allowing best-effort traffic to use the passive reservations is a waste of network resources from the QoS traffic's point of view. Therefore, putting a restriction on the amount of the advance reservation in a subnet would be beneficial from the perspective of network utilisation. In fact, a combination of advance resource reservation and call admission control (CAC) mechanisms should achieve a better performance on managing network resources [3], [9], [10].

III. A Session-to-Mobility Ratio Based Advance Resource Reservation Scheme

In this section, we propose a session-to-mobility ratio (SMR) based advance resource reservation scheme which achieves a better utilisation of the network resources by balancing the amount of active reservations and passive reservations in the network. The proposed scheme consists of two admission mechanisms: *passive reservation bounding* and *SMR based replacement*. We should point out that we are not designing a complete advance resource reservation signalling protocol but investigating an efficient way of using network resources.

A. Passive Reservation Bounding Mechanism

In the ordinary advance resource reservation, since the passive reservation requests are treated in the same way as the active reservation requests, there is no restriction on the amount of passive reservations. As we discussed in Section II, this type of reservation is a waste of resources from the perspective of QoS traffic, it then follows that it is better to give higher priority to the active reservation requests because this type of requests implies there are QoS sessions that really need the required bandwidth.

We borrow the concept of a *channel*, which is widely used in cellular networks, and reserve part of the network resources only for active reservations. Therefore, the channels (i.e., resources) of the network are partitioned into two types: *Standard Channels* and *Dedicated Channels*. In our scheme, there is an enhanced agent (EA) (e.g. PA and MP) in each subnet and they monitor the network resources and operate as *bandwidth brokers* to admit different types of requests. A standard channel can be used for both active reservation and passive reservation, while a dedicated channel can only be used for active reservation. When all the standard channels are full, the passive reservation requests are blocked. In this way, the number of passive reservations in the subnet is bounded and thus more resources are available for active reservations.

Moreover, in order to avoid over-restricting passive reservations, the EA should allocate the active reservation requests first to the dedicated channels and then to the standard channels when all the dedicated channels are engaged. Therefore, if the total number of channels in a subnet is N and the number of standard channels is S, then the maximum number of passive reservations in the subnet is S and the EA can accept at least $N - S$ active reservation requests.

B. SMR Based Replacement Mechanism

The standard channels of a subnet are scarce resources from the foreign mobile nodes' point of view as they require passive reservations. Therefore, an efficient admission strategy is necessary to determine which foreign mobile node is eligible to acquire a standard channel. Since the essential objective of advance resource reservation is to avoid session dropping caused by the handover of the mobile node, we believe it is better to assign a standard channel to the foreign mobile node who is most likely to handover during a session. (Here we assume all the QoS sessions are of the same type.)

Previously, the frequency of handover of a mobile node is characterised by its session-to-mobility ratio (SMR) which is the ratio of session arrival rate to handover rate [11], [12]. In our scheme, we adopt a modified form of SMR which is defined as the ratio of session duration to the mobile node's residence time in a subnet. We believe this revised form of SMR has a clearer representation of the handover frequency of the mobile node since handover is the behaviour of the mobile node during a session and the session arrival rate has no direct relationship with the session duration.

We assume every EA has the SMR information (by either statistical or history-based prediction algorithms) of the mobile

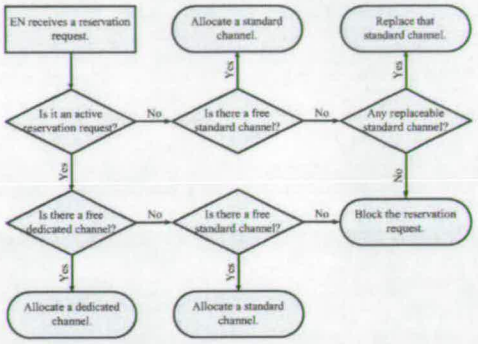


Fig. 1. SMR based advance resource reservation process

nodes in its administrating subnet. The mechanism by which this is achieved is beyond the scope of this work. Our SMR based replacement mechanism works as follows: Assume there is a foreign mobile node which wants to make a passive reservation in the local subnet. The EA of that foreign mobile node will inform the local EA of the SMR of the requesting foreign mobile node. Then

- If there are free standard channels, the local EA will allocate one to the foreign mobile node for passive reservation.
- If there is no free standard channel, the local EA will compare and find out whether the SMR of the requesting foreign mobile node is larger than the smallest of the SMRs of the foreign mobile nodes that have already been allocated standard channels. If it is, the foreign mobile node with the smallest SMR is replaced by the requesting foreign mobile node, i.e., the standard channel (passive reservation path) is re-allocated to the requesting mobile node. Otherwise, the request is blocked.

We should point out that we do not apply the SMR based replacement mechanism to active reservation requests because this would affect the ongoing QoS sessions. On the other hand, the re-allocation of passive reservation paths has no effect on the QoS sessions of the foreign mobile nodes since they are not actively using them.

C. SMR Based Advance Resource Reservation Scheme

In the following we describe the operation of the SMR based advance resource reservation scheme from an EA's perspective. Fig. 1 shows the allocation procedure of our scheme. An EA receives two types of reservation requests.

- **For the active reservation requests:** When a local mobile node requires an active reservation, the local EA will allocate a free channel to the mobile node (a dedicated channel is chosen first, or if one is not available, then a standard channel). When the mobile node finishes the session, the reserved channel is released. However, unlike the other proposed schemes, the active reservation is also released if the local mobile node hands over out of the local subnet and instead the local EA receives a passive reservation request from that mobile node. (In ordinary schemes the active reservation is switched to passive reservation for that mobile node.) We do this to guarantee that the number of passive reservation in a subnet is bounded.
- **For the passive reservation requests:** When a foreign mobile node requires a passive reservation, the local EA of the subnet is informed of the SMR of the requesting foreign mobile node (by the foreign EA) and tries to allocate a standard channel to it. The allocation procedure is described in Section III-B. For the foreign mobile node which fails to obtain a passive reservation in the local subnet, it has to request an active reservation when it hands over into the local subnet (i.e., the *handovered* active reservation request).

In brief, the SMR based advance resource reservation scheme is a CAC enhanced scheme for deploying RSVP in the mobile environment. The CAC is carried out by the EA in each subnet by allocating channels according to the type of requests and taking account of mobility characteristics of the mobile nodes. The motivation for integrating the CAC algorithm is to restrict passive reservations in a subnet and only allow the most eligible mobile nodes to make passive reservations. In this work, we assume the QoS sessions are of the same type and so a mobile node is more eligible in the sense that it has larger SMR value. However, in a broader sense, the type of the QoS sessions should also be considered and it is a very important parameter to determine which mobile node is more suitable for making passive reservations. The introducing of session type will be covered in our future research.

Another important issue we should point out is that although the advance resource reservation schemes look similar to the handover prioritised schemes which are used in the cellular networks, they are different majorly in the ways in which resources are reserved. In the handover prioritised schemes, resources of a subnet are reserved for the mobile nodes in the neighbouring subnets and can be used by anyone that hands over into the subnet. However, in the advance resource reservations schemes, resources are exclusively reserved and therefore the network resource utilisation is deteriorated. In fact, the handover prioritised schemes can be deployed in conjunction with the SMR based advance resource reservation so that the *handovered* active reservation requests are given higher priority.

IV. Performance Evaluation

To compare the performance of the basic and the SMR based advance resource reservation schemes, we build continuous time Markov chain (CTMC) based analytical models using a formal performance modelling formalism named Performance Evaluation Process Algebra (PEPA) [13]. Due to space limitations, we do not present the PEPA models and only give the results in this paper. The details of the PEPA models and parameter settings can be found in [14].

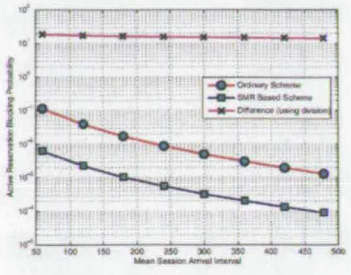


Fig. 2. Active reservation blocking probability vs. session arrival interval

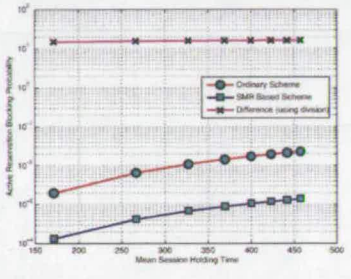


Fig. 3. Active reservation blocking probability vs. session holding time

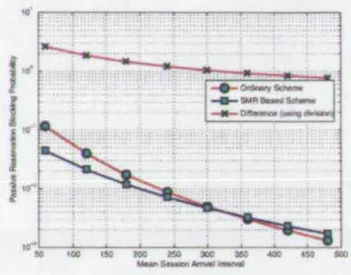


Fig. 4. Passive reservation blocking probability vs. session arrival interval

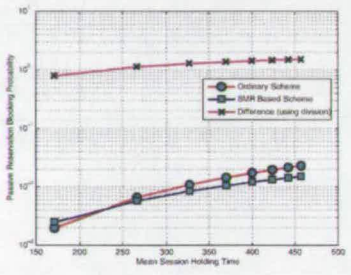


Fig. 5. Passive reservation blocking probability vs. session holding time

Similar to the other literature that focuses on resource management, we investigate the effects of traffic intensity on the blocking probabilities of active and passive reservation requests. These performance metrics reflect the network congestion level for different types of reservations. The traffic intensity can be tuned by two parameters, i.e., the session arrival interval and the session holding time.

A. Active Reservation Blocking Probability

Fig. 2 shows the impact of session arrival interval on the active reservation request blocking probability for both schemes. The results are calculated with the mean session holding time set to 400 units, and the y-axis is in the logarithmic scale. From the figure, we can see that the blocking probability decreases when the interval between sessions gets longer for both schemes. However, the blocking probability in the SMR based scheme is much lower than that in the ordinary scheme. This is because the SMR based scheme sets aside dedicated channels for active reservations. To make the comparison clearer, we calculate the difference of the two schemes using division (the ordinary scheme divided by the SMR based scheme). The plot decreases slightly (even in the logarithmic scale) at larger session arrival intervals, which implies that the SMR based scheme performs better when the traffic intensity is higher (smaller session arrival interval). A similar improvement on the active reservation blocking probability in the SMR based scheme can also be observed in Fig. 3, where the mean session arrival interval is 180 units and the session holding time is the tuning parameter. However, the difference between the two schemes is almost constant and not very sensitive to the session holding time.

B. Passive Reservation Blocking Probability

As we presented in Section III, in the SMR based advance resource reservation scheme, the passive reservations in a subnet are restricted. Therefore, we investigate how passive reservation requests are affected by this restriction. The results are shown in Figs. 4 and 5. An interesting observation from Fig. 4 is that the probability of blocking passive reservation requests in the SMR based scheme is smaller than that in the ordinary scheme when the traffic intensity is high. The reason for this is that in the SMR based scheme, not all the foreign mobile nodes are allowed to make passive reservations and therefore competition for the resources is less severe, i.e., more resources are available to the local mobile nodes and the foreign mobile nodes with large SMR. However, the difference between the two schemes decreases and at session arrival intervals larger than 350 units, the SMR based scheme exceeds its competitor on the passive reservation blocking probability, which is due to the bounded resources for passive reservations. In Fig. 5, we can still identify the advantage of the SMR based scheme when the traffic intensity is high (large session holding

time) and the two schemes produce the same results when the average session holding time is around 230 units.

C. Discussion

By restricting some of the foreign mobile nodes from making passive reservations, the SMR based advance resource reservation scheme achieves a better utilisation of the network resources and reduces both active and passive reservation blocking probability effectively. Although this performance improvement is achieved at the expense of introducing handover interruption to the slow mobile nodes, the SMR based scheme is still reasonable since:

1) Passive reservation blocking only means that a foreign mobile node cannot make advance reservation in the local subnet. Although there could be an interruption when this foreign mobile node hands over into the adjacent subnet, this restriction has no effect on its current QoS session. On the other hand, an active reservation request implies that there is a local mobile node which really needs the requested resources to start its QoS session. Therefore, it is practical to give a higher priority to the active reservation.
2) When a foreign mobile node that is not granted a passive reservation hands over into the local subnet, it becomes a local mobile node and requires an active reservation. Since our scheme sets aside dedicated channels for active reservations and reduces the active reservation blocking probability dramatically, this new local mobile node is highly likely to acquire an active reservation and continue its QoS session. The only expense in this situation is the reservation re-establishment delay. However, the reservation re-establishment delay can be further reduced when RSVP is integrated with micro-mobility management mechanisms [15].
3) The passive reservation brings no revenue while admitting active reservation requests does, therefore we believe restricting the passive reservations in a subnet is beneficial from the network operator's perspective.

V. Conclusion

In this paper, we propose a session-to-mobility ratio (SMR) based advance resource reservation scheme. This scheme aims to balance the amount of active reservations and passive reservations in a subnet. Our motivation is that the passive reservations in a subnet are not actively used by their reservers and thus they waste the network resources from the perspective of the QoS traffic. Our results show that the SMR based scheme can efficiently reduce both active and passive reservation blocking probabilities, which is achieved by setting aside dedicated channels for active reservations and only allowing some of the foreign mobile nodes to make passive reservations in a subnet. Although some of the foreign mobile nodes have to re-establish the reservations when they handover into the subnet, with the help of mobility-aware RSVP, the negative effects of re-establishment can be limited.

Acknowledgement

The work reported in this paper has formed part of the Ubiquitous Services Core Research Programme of the Virtual Centre of Excellence in Mobile & Personal Communications, Mobile VCE, www.mobilevce.com. This research has been funded by the Industrial Companies who are Members of Mobile VCE, with additional financial support from the UK Government's Technology Strategy Board (previously DTI). Fully detailed technical reports on this research are available to Industrial Members of Mobile VCE.

J. Hillston is also supported by an EPSRC Advanced Research Fellowship EP/c543696/01 and EU FET-IST Global Computing 2 project SENSORIA (Software Engineering for Service-Oriented Overlay Computers (IST-3-016004-IP-09)). H. Wang and D. Laurenson acknowledge the support of the Scottish Funding Council for the Joint Research Institute with the Heriot-Watt University which is a part of the Edinburgh Research Partnership.

References

[1] R. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: an Overview," RFC 1633, Jun. 1994.
[2] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification," RFC 2205, Sep. 1997.
[3] S.-J. Leu and R.-S. Chang, "Integrated service mobile Internet: RSVP over mobile IPv4&6," *Mob. Netw. Appl.*, vol. 8, no. 6, pp. 635–642, 2003.
[4] A. K. Talukdar, B. R. Badrinath, and A. Acharya, "MRSVP: a resource reservation protocol for an integrated services network with mobile hosts," *Wirel. Netw.*, vol. 7, no. 1, pp. 5–19, 2001.
[5] C.-C. Tseng, G.-C. Lee, R.-S. Liu, and T.-P. Wang, "HMRSVP: a hierarchical mobile RSVP protocol," *Wirel. Netw.*, vol. 9, no. 2, pp. 95–102, 2003.
[6] W.-T. Chen and L.-C. Huang, "RSVP mobility support: a signaling protocol for integrated services Internet with mobile hosts," in *Proc. International Conference on Computer Communications '00*, 2000, pp. 1283–1292.
[7] N.-F. Huang and W.-E. Chen, "RSVP extensions for real-time services in hierarchical mobile IPv6," *Mob. Netw. Appl.*, vol. 8, no. 6, pp. 625–634, 2003.
[8] J. Zhang, J. Mark, and X. Shen, "A novel resource reservation scheme for handoff in CDMA wireless cellular networks," in *Proc. International Conference on Wireless Communications and Networking '03*, 2003, pp. 2069–2074.
[9] J. Manner, A. L. Toledo, A. Mihailovic, H. L. V. Munoz, E. Hepworth, and Y. Khouaja, "Evaluation of mobility and quality of service interaction," *Comput. Netw.*, vol. 38, no. 2, pp. 137–163, 2002.
[10] G. Huston, "Next Steps for the IP QoS Architecture," RFC 2990, Nov. 2000.
[11] S. Pack, X. S. Shen, J. W. Mark, and J. Pan, "Adaptive route optimization in hierarchical mobile IPv6 networks," *IEEE Trans. Mobile Comput.*, vol. 6, no. 8, pp. 903–914, 2007.
[12] S. Pack, T. Kwon, and Y. Choi, "A mobility-based load control scheme at mobility anchor point in hierarchical mobile IPv6 networks," in *Proc. Global Telecommunications Conference '04*, 2004, pp. 3431–3435.
[13] J. Hillston, *A Compositional Approach to Performance Modelling*. Cambridge University Press, 1996.
[14] H. Wang, D. I. Laurenson, and J. Hillston. (to submit to IEEE Trans. Mobile Comput.) "An SMR based advance resource reservation scheme for deploying RSVP in the mobile environment". [Online]. Available: http://www.see.ed.ac.uk/~s0454305/research.html
[15] H. Wang, D. Laurenson, and J. Hillston, "Evaluation of RSVP and mobility-aware RSVP using performance evaluation process algebra," in *Proc. International Conference on Communications '08*, 2008.

PEPA Analysis of MAP Effects in Hierarchical Mobile IPv6

Hao Wang and Dave Laurenson
Institute for Digital Communications,
Joint Research Institute for Signal & Image Processing,
School of Engineering & Electronics,
University of Edinburgh
Email: {H.Wang, Dave.Laurenson}@ed.ac.uk

Jane Hillston
Laboratory for Foundations of Computer Science,
School of Informatics,
University of Edinburgh
Email: Jane.Hillston@ed.ac.uk

***Abstract*—To overcome the drawbacks of the Mobile IPv6 protocol on handling local mobility management, IETF proposed the HMIPv6 protocol which introduces an intermediate mobility anchor point (MAP) to hide the movement of a mobile node within a local area. However, the MAP forms a bottleneck in the network since all the traffic destined for its served nodes has to go through it. Most research on HMIPv6 focuses on protocol optimisation, and performance analysis of HMIPv6 is usually simulation-based. In this paper, we employ a performance evaluation formalism named PEPA to investigate the performance tradeoffs of MAPs in HMIPv6. Performance measures such as response time and MAP utilisation are presented.**

I. Introduction

To provide continuous connectivity when mobile users change their points of attachment to the Internet, the IETF proposed mobility management protocols Mobile IPv4 [1] and Mobile IPv6 [2] to support global mobility in IP-based networks. In Mobile IPv6-aware networks, a mobile node is always addressable at its home address regardless of its location. Whenever a mobile node moves into a new access network, it acquires one or more care-of addresses representing its current network attachment. The mobile node needs to send Binding Update messages (BUs) which associate its home address with its current care-of address to the mobile node's home agent (HA) and all the correspondent nodes (CNs) it is communicating with. The movement of the mobile node can then be made transparent to the transport and higher-layer by mapping home address to care-of address at the network layer. However, although the Mobile IPv6 protocol supports a route optimisation communication mode, the quality of service will decrease if the mobile node changes its point of attachment so frequently that handoff latency and signalling load caused by Binding Update messages become significant.

To overcome this drawback of the global mobility management protocols, IETF proposed local mobility management protocols such as Cellular IP [3] and Hierarchical Mobile IPv6 (HMIPv6) [4]. The HMIPv6 minimises the amount of signalling outside a local domain by using a new mobility agent, called a Mobility Anchor Point (MAP), that can hide the movement of the mobile node within a local domain. However, the MAP has to operate as a relay node between the mobile node and the CNs since by design all the traffic must go through the MAP. Under heavy traffic conditions, this local mobility management results in the MAPs becoming the bottlenecks of the network and thus network performance is degraded.

In this paper we use a performance evaluation formalism named PEPA to investigate the effects of MAPs on the response time and MAP utilisation in HMIPv6 with a client-server architecture. In particular we investigate the number and placement of MAP nodes within an access network. The rest of paper is organised as follows. In Section II we introduce the PEPA formalism. The HMIPv6 protocol is reviewed in Section III. We present our PEPA model of HMIPv6 and derive performance measures in Sections IV and V respectively. Section VI presents our conclusion.

II. PEPA

Performance Evaluation Process Algebra (PEPA) [5] is both a timed and stochastic extension of classical process algebra such as CCS [6] and CPS [7]. In PEPA a system is described as a component or a group of components that engage in activities. Generally, components model the physical or logical elements of a system and activities characterise the behaviour of these components. Each activity a in PEPA is defined as a pair (α, r) — action type α and activity rate r. The action type can be regarded as the name of the activity and the rate specifies the duration of the activity which is an exponentially distributed random variable. If a component P behaves as Q after completing activity a, then we can denote this transition as:

$$P \xrightarrow{\alpha} Q \text{ or } P \xrightarrow{(\alpha,r)} Q$$

The PEPA formalism provides a small set of operators which are able to express the individual activities of components as well as the interactions between them. We only present the operators we used in our model in this section. For more details about PEPA operators, see [5].

Prefix: $(\alpha, r).P$

The component $(\alpha, r).P$ carries out an activity that is of action type α and has a delay that is exponentially distributed with rate r, which gives an average delay of $1/r$. After

338

completing this activity, the component $(\alpha, r)\,.P$ behaves as component P.

Choice: $P + Q$

The component $P + Q$ may either behave as P or Q. All the enabled activities in P and Q are also enabled in this component and compete with each other. The first activity to be completed will be an activity of P or Q and this will distinguish which component wins the race. When the first activity is completed, all the other activities will be abandoned.

Cooperation: $P \underset{L}{\bowtie} Q$

The component $P \underset{L}{\bowtie} Q$ models the interaction between P and Q. The letter L denotes a set of action types that must be carried out by P and Q together. For all activities whose action type is included in L, P and Q must cooperate to complete it. However, P and Q can carry out other activities independently.

Parallel: $P \| Q$

The component $P \| Q$ represents two concurrent but completely independent components. It is shorthand notation for $P \underset{\emptyset}{\bowtie} Q$.

Constant: $P \stackrel{def}{=} Q$

This expression is used to assign names to components. Such expressions may be mutually recursive leading to infinite behaviours over finite states.

Since the duration of the transition in PEPA is exponentially distributed, it has been shown that the stochastic process underlying a PEPA model is a discrete state space, continuous time Markov chain (CTMC). By deriving the steady state probability distribution for the Markov chain, together with the Markov reward models [8], we can achieve performance measures such as utilisation and throughput. Moreover, measures such as response time can also be calculated by transient analysis. These measures can facilitate model verification and system optimisation.

III. Hierarchical Mobile IPv6 Overview

In Hierarchical Mobile IPv6, there is a mobility agent called a Mobility Anchor Point (MAP) that covers a group of access routers (ARs). Each of these ARs represents a different IP access network and a MAP forms a local network domain. Every time a mobile node moves into a MAP domain, it acquires an on-link care-of address (LCoA) referring to the AR to which it is connected and a regional care-of address (RCoA) referring to the MAP domain. Outside the MAP domain, the mobile node is identified by its RCoA and all the packets addressed at RCoA are intercepted by the MAP and forwarded to the mobile node at LCoA. When the mobile node performs a localised handoff, i.e. switches to a different AR within a MAP domain, it just needs to send a local BU to the MAP to change the mapping between its LCoA and RCoA. The mobile node only needs to send BUs to its HA and CNs informing them of its new RCoA when it moves to a new MAP domain. Therefore, the MAP is able to hide the movement of a mobile node within a local network domain, thereby minimising the handoff latency and outbound signalling load.

However, this mobility management scheme requires all the traffic between the mobile nodes and the CNs to go through the MAPs, which can result in them being bottlenecks in the network and thus degrade network utilisation. Most research on HMIPv6 focuses on performance metrics associated with mobility such as signalling cost and handoff latency, etc., and the analysis is usually simulation-based. In this work, we investigate some other valuable metrics related to behaviour between handoffs i.e. the response time of ARs and the utilisation of the MAP.

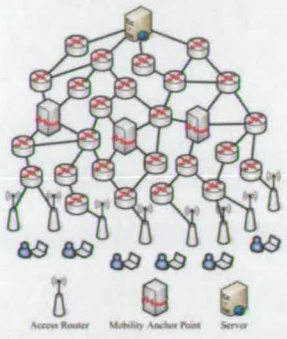


Fig. 1. A Typical Network Architecture of Hierarchical Mobile IPv6

When dealing with networked systems, researchers will typically use simulation tools such as *ns-2* [9] to evaluate performance. In order to avoid problems caused by topologies exhibiting uncharacteristic behaviours, a large set of topologies need to be evaluated, and each one multiple times with different instantiations of traffic flows. Whilst this may be possible for small sized networks, to evaluate large networks, such as those supporting HMIPv6, simulation is not a reliable means of determining performance characteristics. Instead, an analytical approach that is mathematically tractable, but captures the essential characteristics of the network is required.

In this work we present a CTMC-based performance model which is constructed using PEPA to evaluate this local mobility management mechanism. Although there are other Markovian-based performance modelling techniques, such as queueing networks and stochastic Petri nets, PEPA is chosen because its component structure has a better reflection of the system structure, thereby providing a clear description of the system it models. Moreover, since PEPA models can be solved numerically, some restrictions which other modelling approaches must follow to exhibit a product form solution do not constrain PEPA models. In the following sections, this PEPA model of HMIPv6 is described and the effects of the MAPs are analysed.

IV. PEPA Model of HMIPv6

A typical network topology that might be encountered by Hierarchical Mobile IPv6 devices is shown in Fig. 1. The significant elements in the network are the access routers, which act as access points for the mobile devices, mobility anchor points to implement the local mobility function, and the server which serves data destined for the mobile user.

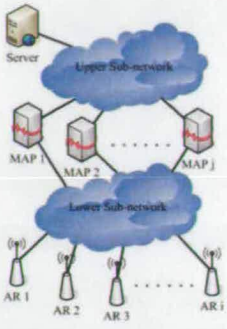


Fig. 2. An Abstract Network Architecture of Hierarchical Mobile IPv6

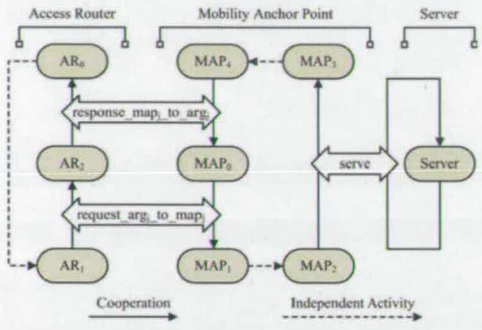


Fig. 3. Cooperations between AR, MAP and Server

The other entities in the figure act as routers for the traffic between these key elements. This network architecture can be considered as the common client-server architecture, where requests and responses have to go through the intermediate MAPs. With any Markov chain representation of interacting entities there is a disproportionate increase in the number of states as the number of entities increases. In order to produce a model that is analytically tractable, and to ensure that the results are more generic than those for a specific topology, the network has been abstracted into the one in Fig. 2. The upper and lower sub-networks model the effect of the entities between the server and MAPs, and those between the MAPs and access routers, are simply represented by delays within the model. We should point out that since we assume the sub-networks are not congested and regard the transmission of requests and responses in the networks as delays, it is not necessary to import components that model the upper and lower sub-networks in the model. This level of simplification is sufficient to derive performance measures that are discussed later in this paper. Moreover, the proposed model is not intended to model the exact HMIPv6 protocol. Instead, it is intended to reflect the interactions between the elementary components of the protocol. Below we show the PEPA definitions for the three types of components.

Access Router (AR): The traffic scenario used in our model is mobile users asking for access of web pages stored on the server through the ARs. An AR covers a number of mobile users and receives requests from them. To investigate the effects of the MAP on one mobile user and simplify the model, we model the ARs as the sources of requests instead of the mobile users and the AR generates the requests at the rate of λ_i. The AR then sends out the requests and waits for the responses. For the ARs that are within the domain of MAP_j, they are considered as an access router group ARG_j. The transmission delays between ARG_j and MAP_j are implemented in cooperations $(request\_arg_j\_to\_map_j, \top)$ and $(response\_map_j\_to\_arg_j, \top)$, corresponding to sending requests and receiving responses (The symbol $\top$ is used in PEPA to denote the passive activity whose activity rate is determined by its cooperation partner). Therefore, the ARs within the same MAP domain have identical behaviour. The component AR is defined as:

$$\begin{aligned} AR_{i0} &\stackrel{def}{=} (ar_i\_request, \lambda_i).AR_{i1}; \\ AR_{i1} &\stackrel{def}{=} (request\_arg_j\_to\_map_j, \top).AR_{i2}; \\ AR_{i2} &\stackrel{def}{=} (response\_map_j\_to\_arg_j, \top).AR_{i0}; \end{aligned}$$

Mobility Anchor Point (MAP): Each MAP handles a certain number of ARs and relays all the traffic between the ARs and the server. The MAP receives requests from the ARs and forwards them to the server. Once the requests are answered at the server, i.e., completing the $(serve, \top)$ cooperation, the MAP collects and sends the responses back to the ARs. The delay of the traffic in both directions is divided into two parts, i.e., the delays in the lower sub-network and upper sub-network. The length of delays in lower and upper sub-networks is determined by the parameters α_{j1} and α_{j2} respectively. Unlike the delay in the lower sub-network, the delay in the higher sub-network is not implemented as a cooperation between the MAPs and the Server since doing this will prohibit the Server from serving other MAPs. It should be pointed out that since we consider the MAPs as the bottlenecks of the network, a MAP is modelled as a scarce resource and is not able to accept requests from other ARs while it is already engaged by one AR. The definition of MAP is given below:

$$\begin{aligned} MAP_{j0} &\stackrel{def}{=} (request\_arg_j\_to\_map_j, \alpha_{j1}).MAP_{j1}; \\ MAP_{j1} &\stackrel{def}{=} (request\_map_j\_to\_server, \alpha_{j2}).MAP_{j2}; \\ MAP_{j2} &\stackrel{def}{=} (serve, \top).MAP_{j3}; \\ MAP_{j3} &\stackrel{def}{=} (response\_server\_to\_map_j, \alpha_{j2}).MAP_{j4}; \\ MAP_{j4} &\stackrel{def}{=} (response\_map_j\_to\_arg_j, \alpha_{j1}).MAP_{j0}; \end{aligned}$$

Server: We assume the server has infinite buffer size and is able to manage as many requests as the MAPs can submit, i.e., it carries out the serve activity in an iterative way. Since the server can cooperate with whichever MAP at a time, it is regarded as a server with random order service strategy. The component Server is defined as:

$$Server \stackrel{def}{=} (serve, \mu).Server;$$

TABLE I
PARAMETERS VALUES

Type (Role)	Average Time (ms)	Rate (1/ms)
$ar_i\_request$ (request rate)	10	0.1
$request\_arg_j\_to\_map_j$ (delay of request in lower sub-network)	5	0.2
$request\_map_j\_to\_server$ (delay of request in upper sub-network)	5	0.2
$response\_server\_to\_map_j$ (delay of response in upper sub-network)	5	0.2
$response\_map_j\_to\_arg_j$ (delay of response in lower sub-network)	5	0.2
serve (service rate)	0.01	100

TABLE II
MAPPING BETWEEN ACCESS ROUTERS AND MAPS IN SCENARIOS I –V

Scenario	MAP1	MAP2	MAP3
I	1,2,3,4,5,6	N.A.	N.A.
II	1,2,3	4,5,6	N.A.
III	1,2,3,4	5,6	N.A.
IV	1,2	3,4	5,6
V	1,2,3	4	5,6

System Definition: This definition specifies how the system is constructed from the defined components. Generally, a PEPA system is defined as the interactions between the components. For our model if there are m ARs and n MAPs then the system is expressed as:

$$System \stackrel{def}{=} (AR_1 \| \cdots \| AR_m) \underset{L_1}{\bowtie} (MAP_1 \| \cdots \| MAP_n) \underset{L_2}{\bowtie} (Server)$$

where

$$L_1 = \{request\_arg_j\_to\_map_j, response\_map_j\_to\_arg_j\},$$

$$L_2 = \{serve\}.$$

The cooperations between the defined components are represented diagrammatically in Fig. 3.

Parameters Setting: To derive performance measures we first need to set the activity rates within the model. There are three types of activity in the model, i.e. the request rate of the AR, the delay in the sub-networks and the service rate. In our experiments, we assume an AR receives 10^2 web page requests per second from the mobile users and the server is able to handle 10^5 requests every second. For the delay of the sub-networks, we assume there are two or three hops for every message transmission, and based on the default value used in *ns-2*, the delay of each hop lies between 1 and 2ms. Therefore we set the average delay of the sub-networks to be 5ms. These activity rates are shown in Table I. Moreover, we assume the MAPs sit in the middle of the network and both requests and responses take the same routes, hence the activity rates for delay in the sub-networks are set to be the same.

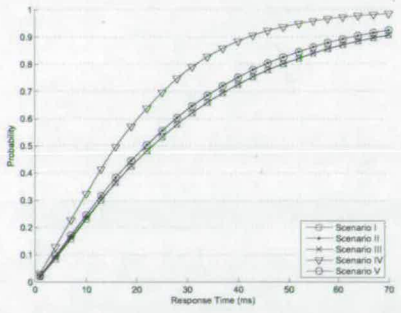


Fig. 4. CDF of the Response Time of AR1 in Different Scenarios

V. PERFORMANCE EVALUATION

Unlike most of the performance analysis of HMIPv6 that focus on handoff delay and signalling cost, the performance measures we investigate are the response time of the ARs and the utilisation of the MAPs. We carry out our experiments using the PEPA Workbench [10] and associated tools such as ipc [11] and Hydra [12]. More details on these tools can be found at http://www.dcs.ed.ac.uk/pepa. We analyse five network scenarios with 6 ARs and different numbers of MAPs. The connectivity of the ARs and the MAPs are shown in Table II.

A. Response Time

Response time is the time between an AR sending a request to the AR receiving the response. The response time for an AR in our model comprises three time periods, namely: queueing time at the MAP; delay in the sub-networks; and waiting time at the server.

We first investigate the response time, in the form of a cumulative distribution function (CDF), of AR1 in all of the five network scenarios. The result is shown in Fig. 4. Given the request rates of 0.1 for all ARs, the response time of AR1 degrades and reaches the limit rapidly as the MAP domain size increases. As we can see from the figure, forcing MAP1 to serve 4 ARs is as bad as connecting 6 ARs to it, and about 10% of the requests cannot be answered within 70ms. In the scenarios where the MAP1 serves three ARs, since the MAPs behave independently and do not block the server, AR1 has the same response time distribution, which is shorter but the improvement is not significant. However, making AR1 compete with only AR2 for MAP1 provides us a much faster response from the server. This suggests that the response time of an AR strongly relies on the domain size of its MAP and more MAPs does not necessarily mean a faster response.

The relationship between the response time and request rate is another performance metric we are interested in. We examine the response time of AR1 against the request rate of AR2 in scenario IV, where AR1 has the shortest response time.

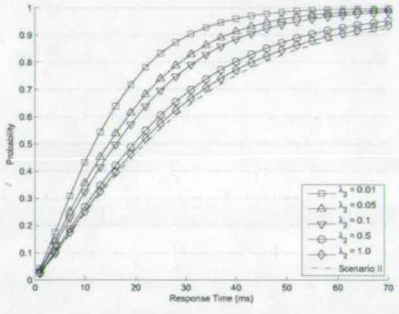


Fig. 5. CDF of the Response Time of AR1 in Scenario IV with Increasing Request Rate at AR2

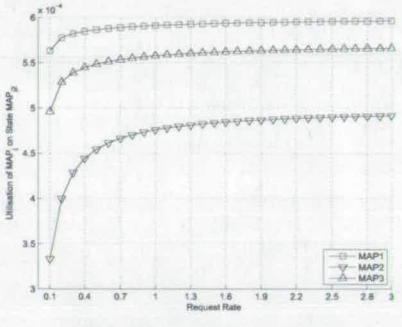


Fig. 6. Utilisation of MAP_j on State MAP_{j2} with Increasing Request Rate at All ARs in Scenario V

The result is shown in Fig. 5. It is clear to see that AR1 has to wait a longer time if the request rate of AR2 increases. This is because as AR2 sends requests more frequently, MAP1 is more likely to be engaged by AR2, which prevents AR1 from using MAP1. Comparing Fig. 4 and Fig. 5, it can be found that as we increase the request rate of AR2, the response time of AR1 is getting close to that in scenario II where three ARs are attached to MAP1. This means that the heavily loaded AR2 is so greedy that it starves AR1 of MAP1 as if another AR were introduced. On the other hand, AR1 could receive faster response if its competitor AR2 is relatively lightly loaded with request rate much smaller than 0.1. Therefore, it can be concluded that the response time of an access router also relies on the workload of its MAP and a heavily loaded AR is best connected to a MAP with other ARs which are lightly loaded.

B. MAP Utilisation

Another important performance measure is the utilisation of each MAP. This metric can tell us whether a MAP is well or underutilised. In [5] the utilisation is defined as the fraction of the time in which a component stays in different states.

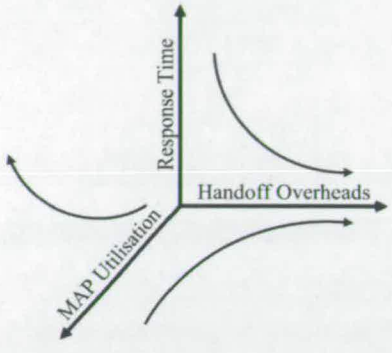


Fig. 7. Trade-off Between Response Time, MAP Utilisation and Handoff Overheads

Although in practice a MAP does not have states in which it can engage in the activities corresponding to the delay in the network, those states can represent the idle phases of a MAP. To investigate the utilisation of a MAP, we can analyse the proportion of time that each MAP spends on the serve activity, i.e., being in the MAP_{j2} state, which can indicate the workload of that MAP.

We carry out the experiments using scenario V and increasing request rates of all ARs from 0.1 to 3.0 and keeping the other activity rates the same as set in Table I. The results are shown in Fig. 6.

As the request rates increase, the MAPs can engage in the service activity more frequently. However, the speed of increase is different in each MAP. This can be explained as follows: In our model, the mean sojourn time for MAP_j in state MAP_{j0} is comprised of two periods, namely, waiting for a request and delay of a request. Since the mean waiting time (for a request) of a MAP is inversely proportional to the request rate and the delay in the lower sub-network is not affected by the request rate, the mean sojourn time in state MAP_{j0} does not change in a linear way, which results in a non-linear increase in sojourn time in all the other states, including MAP_{j2}. From Fig. 6, we can find that even if the request rate of AR4 rises to 3.0, the utilisation of MAP2 is very close to that of MAP3 where both of its ARs have requests rate of 0.1. Also, MAP1 can have almost the same utilisation as MAP2 with very busy ARs while its ARs are only lightly loaded. This means that connecting a heavy load to a MAP does not necessarily mean a better utilisation of that MAP, but a larger MAP domain size can improve its utilisation.

VI. Conclusion

In this paper we have presented the performance evaluation formalism PEPA and demonstrated its application to the HMIPv6 protocol with a common client-server network architecture. Since a MAP is a bottleneck in the network, performance modelling can play an essential role in assessing its effects on the network. We investigate the impacts of the

342

MAP on the response time of AR and the utilisation of MAP. The response time is a very important metric because it is a measure of the network QoS that is experienced by the user. A large response time usually results in unsatisfactory service or even service interruption. The results indicate that the MAP domain size is an important performance factor to the response time of ARs. A smaller MAP domain size provides a better response time. However, reducing the MAP domain size implies uneconomical network deployment with more MAP nodes and more inter MAP domain handoff, which minimise the expected benefits of HMIPv6. Moreover, the MAP domain size also affects the utilisation of a MAP, and larger domain size implies a better use of that MAP. This kind of trade-off is shown in Fig 7. As we increase the MAP domain, we can achieve better MAP utilisation and smaller handoff overheads, at the expense of larger response time.

Furthermore, the results also show that a heavily loaded AR can starve the other ARs sharing the same MAP. An intuitive solution of this problem would be connecting the heavily loaded AR to a MAP with light workload. However, this requirement cannot be easily fulfilled in mobile communication scenarios where the heavily loaded ARs are continually changing. This issue is important if we want to integrate the mobility management and QoS mechanisms. In the mobile network (NEMO) scenario [13], where mobile nodes move as a group, this can easily reduce the QoS of their visited AR. This situation should be improved if ARs can choose their MAPs adaptively according to their requested load. In our future work, we will design a more sophisticated PEPA model of HMIPv6 that can express different types of data traffic and investigate possible mechanisms of integrating mobility and QoS management.

ACKNOWLEDGEMENT

The work reported in this paper has formed part of the Ubiquitous Services Core Research Programme of the Virtual Centre of Excellence in Mobile & Personal Communications, Mobile VCE, www.mobilevce.com. This research has been funded by the DTI-led Technology Programme and by the Industrial Companies who are Members of Mobile VCE. Fully detailed technical reports on this research are available to Industrial Members of Mobile VCE. J. Hillston is also supported by EPSRC Advanced Research Fellowship EP/c543696/01 and EU FET-IST Global Computing 2 project SENSORIA (Software Engineering for Service-Oriented Overlay Computers (IST-3-016004-IP-09)). H. Wang and D. Laurenson acknowledge the support of the Scottish Funding Council for the Joint Research Institute with the Heriot-Watt University which is a part of the Edinburgh Research Partnership.

REFERENCES

[1] C. Perkins, "IP Mobility Support," RFC 2002, Oct. 1999.
[2] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6," RFC 3775, Jun. 2004.
[3] A. T. Campbell, J. Gomez, and A. G. Valko, "An overview of cellular ip," in *Proc. Wireless Communications and Networking Conference '99*, Sep. 1999, pp. 606–610.
[4] H. Soliman, C. Castelluccia, K. E. Malki, and L. Bellier, "Hierarchical Mobile IPv6 Mobility Management (HMIPv6)," RFC 4140, Aug. 2005.
[5] J. Hillston, *A Compositional Approach to Performance Modelling*. Cambridge University Press, 1996.
[6] R. Milner, *Communication and Concurrency*. Prentice Hall, 1989.
[7] C. A. R. Hoare, *Communicating Sequential Processes*. Prentice Hall, 1985.
[8] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. John Wiley & Sons, 1998.
[9] The Network Simulator - *ns*-2. [Online]. Available: http://www.isi.edu/nsnam/ns/
[10] S. Gilmore and J. Hillston, "The PEPA workbench: A tool to support a process algebra-based approach to performance modelling," in *Proc. Modelling Techniques and Tools for Computer Performance Evaluation '94*, May 1994, pp. 353–368.
[11] J. Bradley, N. Dingle, S. Gilmore, and W. Knottenbelt, "Derivation of passage-time densities in PEPA models using IPC: The Imperial PEPA Compiler," in *Proc. Modeling, Analysis and Simulation of Computer and Telecommunications Systems '03*, Oct. 2003, pp. 344–351.
[12] J. Bradley, N. Dingle, S. Gilmore, and W. Knottenbelt, "Extracting passage times from PEPA models with the HYDRA tool: A case study," in *Proc. UK Performance Engineering Workshop '03*, Jul. 2003, pp. 79–90.
[13] V. Devarapalli, R. Wakikawa, A. Petrescu, and P. Thubert, "Network Mobility (NEMO) Basic Support Protocol," RFC 3963, Jan. 2005.